Robust Principal Component Analysis?

Emmanuel J. Candès[1,2], Xiaodong Li[2], Yi Ma[3,4], and John Wright[4]

[1] Department of Statistics, Stanford University, Stanford, CA 94305

[2] Department of Mathematics, Stanford University, Stanford, CA 94305

[3,4] Electrical and Computer Engineering, UIUC, Urbana, IL 61801

[4] Microsoft Research Asia, Beijing, China

December 17, 2009

Abstract

This paper is about a curious phenomenon. Suppose we have a data matrix, which is the superposition of a low-rank component and a sparse component. Can we recover each component individually? We prove that under some suitable assumptions, it is possible to recover both the low-rank and the sparse components *exactly* by solving a very convenient convex program called *Principal Component Pursuit*; among all feasible decompositions, simply minimize a weighted combination of the nuclear norm and of the ℓ_1 norm. This suggests the possibility of a principled approach to robust principal component analysis since our methodology and results assert that one can recover the principal components of a data matrix even though a positive fraction of its entries are arbitrarily corrupted. This extends to the situation where a fraction of the entries are missing as well. We discuss an algorithm for solving this optimization problem, and present applications in the area of video surveillance, where our methodology allows for the detection of objects in a cluttered background, and in the area of face recognition, where it offers a principled way of removing shadows and specularities in images of faces.

Keywords. Principal components, robustness vis-a-vis outliers, nuclear-norm minimization, ℓ_1-norm minimization, duality, low-rank matrices, sparsity, video surveillance.

1 Introduction

1.1 Motivation

Suppose we are given a large data matrix M, and know that it may be decomposed as

$$M = L_0 + S_0,$$

where L_0 has low-rank and S_0 is sparse; here, both components are of arbitrary magnitude. We do not know the low-dimensional column and row space of L_0, not even their dimension. Similarly, we do not know the locations of the nonzero entries of S_0, not even how many there are. Can we hope to recover the low-rank and sparse components both accurately (perhaps even exactly) and efficiently?

A *provably correct* and *scalable* solution to the above problem would presumably have an impact on today's data-intensive scientific discovery.[1] The recent explosion of massive amounts of high-

---

[1]Data-intensive computing is advocated by Jim Gray as the fourth paradigm for scientific discovery [24].

dimensional data in science, engineering, and society presents a challenge as well as an opportunity to many areas such as image, video, multimedia processing, web relevancy data analysis, search, biomedical imaging and bioinformatics. In such application domains, data now routinely lie in thousands or even billions of dimensions, with a number of samples sometimes of the same order of magnitude.

To alleviate the curse of dimensionality and scale,[2] we must leverage on the fact that such data have low intrinsic dimensionality, e.g. that they lie on some low-dimensional subspace [15], are sparse in some basis [13], or lie on some low-dimensional manifold [4, 46]. Perhaps the simplest and most useful assumption is that the data all lie near some low-dimensional subspace. More precisely, this says that if we stack all the data points as column vectors of a matrix M, the matrix should have (approximately) low-rank: mathematically,

$$M = L_0 + N_0,$$

where L_0 has low-rank and N_0 is a small perturbation matrix. Classical *Principal Component Analysis* (PCA) [15, 25, 27] seeks the best (in an ℓ^2 sense) rank-k estimate of L_0 by solving

$$\begin{array}{ll} \text{minimize} & \|M - L\| \\ \text{subject to} & \text{rank}(L) \le k. \end{array}$$

(Throughout the paper, $\|M\|$ denotes the 2-norm; that is, the largest singular value of M.) This problem can be efficiently solved via the singular value decomposition (SVD) and enjoys a number of optimality properties when the noise N_0 is small and i.i.d. Gaussian.

Robust PCA. PCA is arguably the most widely used statistical tool for data analysis and dimensionality reduction today. However, its brittleness with respect to *grossly* corrupted observations often puts its validity in jeopardy – a single grossly corrupted entry in M could render the estimated $\hat{L}$ arbitrarily far from the true L_0. Unfortunately, gross errors are now ubiquitous in modern applications such as image processing, web data analysis, and bioinformatics, where some measurements may be arbitrarily corrupted (due to occlusions, malicious tampering, or sensor failures) or simply irrelevant to the low-dimensional structure we seek to identify. A number of natural approaches to robustifying PCA have been explored and proposed in the literature over several decades. The representative approaches include influence function techniques [26, 47], multivariate trimming [19], alternating minimization [28], and random sampling techniques [17]. Unfortunately, none of these existing approaches yields a polynomial-time algorithm with strong performance guarantees under broad conditions[3]. The new problem we consider here can be considered as an idealized version of *Robust PCA*, in which we aim to recover a *low-rank* matrix L_0 from highly corrupted measurements $M = L_0 + S_0$. Unlike the small noise term N_0 in classical PCA, the entries in S_0 can have arbitrarily large magnitude, and their support is assumed to be *sparse* but unknown[4].

---

[2]We refer to either the complexity of algorithms that increases drastically as dimension increases, or to their performance that decreases sharply when scale goes up.

[3]Random sampling approaches guarantee near-optimal estimates, but have complexity exponential in the rank of the matrix L_0. Trimming algorithms have comparatively lower computational complexity, but guarantee only locally optimal solutions.

[4]The unknown support of the errors makes the problem more difficult than the matrix completion problem that has been recently much studied.

Applications. There are many important applications in which the data under study can naturally be modeled as a low-rank plus a sparse contribution. All the statistical applications, in which robust principal components are sought, of course fit our model. Below, we give examples inspired by contemporary challenges in computer science, and note that depending on the applications, either the low-rank component or the sparse component could be the object of interest:

- *Video Surveillance.* Given a sequence of surveillance video frames, we often need to identify activities that stand out from the background. If we stack the video frames as columns of a matrix M, then the low-rank component L_0 naturally corresponds to the stationary background and the sparse component S_0 captures the moving objects in the foreground. However, each image frame has thousands or tens of thousands of pixels, and each video fragment contains hundreds or thousands of frames. It would be impossible to decompose M in such a way unless we have a truly scalable solution to this problem. In Section 4, we will show the results of our algorithm on video decomposition.

- *Face Recognition.* It is well known that images of a convex, Lambertian surface under varying illuminations span a low-dimensional subspace [1]. This fact has been a main reason why low-dimensional models are mostly effective for imagery data. In particular, images of a human's face can be well-approximated by a low-dimensional subspace. Being able to correctly retrieve this subspace is crucial in many applications such as face recognition and alignment. However, realistic face images often suffer from self-shadowing, specularities, or saturations in brightness, which make this a difficult task and subsequently compromise the recognition performance. In Section 4, we will show how our method is able to effectively remove such defects in face images.

- *Latent Semantic Indexing.* Web search engines often need to analyze and index the content of an enormous corpus of documents. A popular scheme is the *Latent Semantic Indexing* (LSI) [14, 43]. The basic idea is to gather a document-versus-term matrix M whose entries typically encode the relevance of a term (or a word) to a document such as the frequency it appears in the document (e.g. the TF/IDF). PCA (or SVD) has traditionally been used to decompose the matrix as a low-rank part plus a residual, which is not necessarily sparse (as we would like). If we were able to decompose M as a sum of a low-rank component L_0 and a sparse component S_0, then L_0 could capture common words used in all the documents while S_0 captures the few key words that best distinguish each document from others.

- *Ranking and Collaborative Filtering.* The problem of anticipating user tastes is gaining increasing importance in online commerce and advertisement. Companies now routinely collect user rankings for various products, e.g., movies, books, games, or web tools, among which the Netflix Prize for movie ranking is the best known [41]. The problem is to use incomplete rankings provided by the users on some of the products to predict the preference of any given user on any of the products. This problem is typically cast as a low-rank matrix completion problem. However, as the data collection process often lacks control or is sometimes even *ad hoc* – a small portion of the available rankings could be noisy and even tampered with. The problem is more challenging since we need to simultaneously complete the matrix and correct the errors. That is, we need to infer a low-rank matrix L_0 from a set of incomplete and corrupted entries. In Section 1.6, we will see how our results can be extended to this situation.

Similar problems also arise in many other applications such as graphical model learning, linear system identification, and coherence decomposition in optical systems, as discussed in [12]. All in all, the new applications we have listed above require solving the low-rank and sparse decomposition problem for matrices of extremely high dimension and under much broader conditions, a goal this paper aims to achieve.

1.2 A surprising message

At first sight, the separation problem seems impossible to solve since the number of unknowns to infer for L_0 and S_0 is twice as many as the given measurements in $M \in \mathbb{R}^{n_1 \times n_2}$. Furthermore, it seems even more daunting that we expect to reliably obtain the low-rank matrix L_0 with errors in S_0 of arbitrarily large magnitude.

In this paper, we are going to see that very surprisingly, not only can this problem be solved, it can be solved by *tractable* convex optimization. Let $\|M\|_* := \sum_i \sigma_i(M)$ denote the nuclear norm of the matrix M, i.e. the sum of the singular values of M, and let $\|M\|_1 = \sum_{ij} |M_{ij}|$ denote the ℓ_1-norm of M seen as a long vector in $\mathbb{R}^{n_1 \times n_2}$. Then we will show that under rather weak assumptions, the *Principal Component Pursuit* (PCP) estimate solving[5]

$$
\begin{aligned}
\text{minimize} \quad & \|L\|_* + \lambda \|S\|_1 \\
\text{subject to} \quad & L + S = M
\end{aligned}
\tag{1.1}
$$

exactly recovers the low-rank L_0 and the sparse S_0. Theoretically, this is guaranteed to work even if the rank of L_0 grows almost linearly in the dimension of the matrix, and the errors in S_0 are up to a constant fraction of all entries. Algorithmically, we will see that the above problem can be solved by efficient and scalable algorithms, at a cost not so much higher than the classical PCA. Empirically, our simulations and experiments suggest this works under surprisingly broad conditions for many types of real data. In Section 1.5, we will comment on the similar approach taken in the paper [12], which was released during the preparation of this manuscript.

1.3 When does separation make sense?

A normal reaction is that the objectives of this paper cannot be met. Indeed, there seems to not be enough information to perfectly disentangle the low-rank and the sparse components. And indeed, there is some truth to this, since there obviously is an identifiability issue. For instance, suppose the matrix M is equal to $e_1 e_1^*$ (this matrix has a one in the top left corner and zeros everywhere else). Then since M is both sparse and low-rank, how can we decide whether it is low-rank or sparse? To make the problem meaningful, we need to impose that the low-rank component L_0 is not sparse. In this paper, we will borrow the general notion of incoherence introduced in [8] for the matrix completion problem; this is an assumption concerning the singular vectors of the low-rank component. Write the singular value decomposition of $L_0 \in \mathbb{R}^{n_1 \times n_2}$ as

$$
L_0 = U\Sigma V^* = \sum_{i=1}^{r} \sigma_i u_i v_i^*,
$$

---

[5]Although the name naturally suggests an emphasis on the recovery of the low-rank component, we reiterate that in some applications, the sparse component truly is the object of interest.

where r is the rank of the matrix, $\sigma_1, \ldots, \sigma_r$ are the positive singular values, and $U = [u_1, \ldots, u_r]$, $V = [v_1, \ldots, v_r]$ are the matrices of left- and right-singular vectors. Then the incoherence condition with parameter μ states that

$$\max_i \|U^* e_i\|^2 \leq \frac{\mu r}{n_1}, \quad \max_i \|V^* e_i\|^2 \leq \frac{\mu r}{n_2}, \tag{1.2}$$

and

$$\|UV^*\|_\infty \leq \sqrt{\frac{\mu r}{n_1 n_2}}. \tag{1.3}$$

Here and below, $\|M\|_\infty = \max_{i,j} |M_{ij}|$, i.e. is the ℓ_∞ norm of M seen as a long vector. Note that since the orthogonal projection P_U onto the column space of U is given by $P_U = UU^*$, (1.2) is equivalent to $\max_i \|P_U e_i\|^2 \leq \mu r/n_1$, and similarly for P_V. As discussed in earlier references [8, 10, 22], the incoherence condition asserts that for small values of μ, the singular vectors are reasonably spread out – in other words, not sparse.

Another identifiability issue arises if the sparse matrix has low-rank. This will occur if, say, all the nonzero entries of S occur in a column or in a few columns. Suppose for instance, that the first column of S_0 is the opposite of that of L_0, and that all the other columns of S_0 vanish. Then it is clear that we would not be able to recover L_0 and S_0 by any method whatsoever since $M = L_0 + S_0$ would have a column space equal to, or included in that of L_0. To avoid such meaningless situations, we will assume that the sparsity pattern of the sparse component is selected uniformly at random.

1.4   Main result

The surprise is that under these minimal assumptions, the simple PCP solution perfectly recovers the low-rank and the sparse components, provided of course that the rank of the low-rank component is not too large, and that the sparse component is reasonably sparse. Below, $n_{(1)} = \max(n_1, n_2)$ and $n_{(2)} = \min(n_1, n_2)$.

Theorem 1.1 *Suppose L_0 is $n \times n$, obeys (1.2)–(1.3), and that the support set of S_0 is uniformly distributed among all sets of cardinality m. Then there is a numerical constant c such that with probability at least $1 - cn^{-10}$ (over the choice of support of S_0), Principal Component Pursuit (1.1) with $\lambda = 1/\sqrt{n}$ is exact, i.e. $\hat{L} = L_0$ and $\hat{S} = S_0$, provided that*

$$\mathrm{rank}(L_0) \leq \rho_r n \, \mu^{-1} (\log n)^{-2} \quad and \quad m \leq \rho_s \, n^2. \tag{1.4}$$

Above, ρ_r and ρ_s are positive numerical constants. In the general rectangular case where L_0 is $n_1 \times n_2$, PCP with $\lambda = 1/\sqrt{n_{(1)}}$ succeeds with probability at least $1 - cn_{(1)}^{-10}$, provided that $\mathrm{rank}(L_0) \leq \rho_r n_{(2)} \, \mu^{-1} (\log n_{(1)})^{-2}$ and $m \leq \rho_s \, n_1 n_2$.

In other words, matrices L_0 whose singular vectors—or principal components—are reasonably spread can be recovered with probability nearly one from arbitrary and completely unknown corruption patterns (as long as these are randomly distributed). In fact, this works for large values of the rank, i.e. on the order of $n/(\log n)^2$ when μ is not too large. We would like to emphasize that the only 'piece of randomness' in our assumptions concerns the locations of the nonzero entries of S_0; everything else is deterministic. In particular, all we require about L_0 is that its singular vectors are not spiky. Also, we make no assumption about the magnitudes or signs of the nonzero

5

entries of S_0. To avoid any ambiguity, our model for S_0 is this: take an *arbitrary* matrix S and set to zero its entries on the random set Ω^c; this gives S_0.

A rather remarkable fact is that there is no tuning parameter in our algorithm. Under the assumption of the theorem, minimizing

$$\|L\|_* + \frac{1}{\sqrt{n_{(1)}}}\|S\|_1, \quad n_{(1)} = \max(n_1, n_2)$$

always returns the correct answer. This is surprising because one might have expected that one would have to choose the right scalar λ to balance the two terms in $\|L\|_* + \lambda\|S\|_1$ appropriately (perhaps depending on their relative size). This is, however, clearly not the case. In this sense, the choice $\lambda = 1/\sqrt{n_{(1)}}$ is universal. Further, it is not a priori very clear why $\lambda = 1/\sqrt{n_{(1)}}$ is a correct choice no matter what L_0 and S_0 are. It is the mathematical analysis which reveals the correctness of this value. In fact, the proof of the theorem gives a whole range of correct values, and we have selected a sufficiently simple value in that range.

Another comment is that one can obtain results with larger probabilities of success, i.e. of the form $1 - O(n^{-\beta})$ (or $1 - O(n_{(1)}^{-\beta})$) for $\beta > 0$ at the expense of reducing the value of ρ_r.

1.5   Connections with prior work and innovations

The last year or two have seen the rapid development of a scientific literature concerned with the *matrix completion* problem introduced in [8], see also [7,10,22,23,29] and the references therein. In a nutshell, the matrix completion problem is that of recovering a low-rank matrix from only a small fraction of its entries, and by extension, from a small number of linear functionals. Although other methods have been proposed [29], the method of choice is to use convex optimization [7,10,22,23,45]: among all the matrices consistent with the data, simply find that with minimum nuclear norm. The papers cited above all prove the mathematical validity of this approach, and our mathematical analysis borrows ideas from this literature, and especially from those pioneered in [8]. Our methods also much rely on the powerful ideas and elegant techniques introduced by David Gross in the context of quantum-state tomography [22,23]. In particular, the clever golfing scheme [22] plays a crucial role in our analysis, and we introduce two novel modifications to this scheme.

Despite these similarities, our ideas depart from the literature on matrix completion on several fronts. First, our results obviously are of a different nature. Second, we could think of our separation problem, and the recovery of the low-rank component, as a matrix completion problem. Indeed, instead of having a fraction of observed entries available and the other missing, we have a fraction available, but do not know which one, while the other is not missing but entirely corrupted altogether. Although, this is a harder problem, one way to think of our algorithm is that it simultaneously detects the corrupted entries, and perfectly fits the low-rank component to the remaining entries that are deemed reliable. In this sense, our methodology and results go beyond matrix completion. Third, we introduce a novel de-randomization argument that allows us to fix the signs of the nonzero entries of the sparse component. We believe that this technique will have many applications. One such application is in the area of compressive sensing, where assumptions about the randomness of the signs of a signal are common, and merely made out of convenience rather than necessity; this is important because assuming independent signal signs may not make much sense for many practical applications when the involved signals can all be non-negative (such as images).

We mentioned earlier the related work [12], which also considers the problem of decomposing a given data matrix into sparse and low-rank components, and gives sufficient conditions for convex programming to succeed. These conditions are phrased in terms of two quantities. The first is the maximum ratio between the ℓ_∞ norm and the operator norm, restricted to the subspace generated by matrices whose row or column spaces agree with those of L_0. The second is the maximum ratio between the operator norm and the ℓ_∞ norm, restricted to the subspace of matrices that vanish off the support of S_0. Chandrasekaran et. al. show that when the product of these two quantities is small, then the recovery is exact for a certain interval of the regularization parameter [12].

One very appealing aspect of this condition is that it is completely deterministic: it does not depend on any random model for L_0 or S_0. It yields a corollary that can be easily compared to our result: suppose $n_1 = n_2 = n$ for simplicity, and let μ_0 be the smallest quantity satisfying (1.2), then correct recovery occurs whenever

$$\max_j \{i : [S_0]_{ij} \neq 0\} \times \sqrt{\mu_0 r/n} < 1/12.$$

The left-hand side is at least as large as $\rho_s \sqrt{\mu_0 n r}$, where ρ_s is the fraction of entries of S_0 that are nonzero. Since $\mu_0 \geq 1$ always, this statement only guarantees recovery if $\rho_s = O((nr)^{-1/2})$; i.e., even when $\text{rank}(L_0) = O(1)$, only vanishing fractions of the entries in S_0 can be nonzero.

In contrast, our result shows that for incoherent L_0, correct recovery occurs with high probability for $\text{rank}(L_0)$ on the order of $n/[\mu \log^2 n]$ and a number of nonzero entries in S_0 on the order of n^2. That is, matrices of large rank can be recovered from non-vanishing fractions of sparse errors. This improvement comes at the expense of introducing one piece of randomness: a uniform model on the error support.[6]

Our analysis has one additional advantage, which is of significant practical importance: it identifies a simple, non-adaptive choice of the regularization parameter λ. In contrast, the conditions on the regularization parameter given by Chandrasekaran et al. depend on quantities which in practice are not known a-priori. The experimental section of [12] suggests searching for the correct λ by solving many convex programs. Our result, on the other hand, demonstrates that the simple choice $\lambda = 1/\sqrt{n}$ works with high probability for recovering any square incoherent matrix.

1.6 Implications for matrix completion from grossly corrupted data

We have seen that our main result asserts that it is possible to recover a low-rank matrix even though a significant fraction of its entries are corrupted. In some applications, however, some of the entries may be missing as well, and this section addresses this situation. Let $\mathcal{P}_\Omega$ be the orthogonal projection onto the linear space of matrices supported on $\Omega \subset [n_1] \times [n_2]$,

$$\mathcal{P}_\Omega X = \begin{cases} X_{ij}, & (i,j) \in \Omega, \\ 0, & (i,j) \notin \Omega. \end{cases}$$

Then imagine we only have available a few entries of $L_0 + S_0$, which we conveniently write as

$$Y = \mathcal{P}_{\Omega_{\text{obs}}}(L_0 + S_0) = \mathcal{P}_{\Omega_{\text{obs}}} L_0 + S'_0;$$

---

[6]Notice that the bound of [12] depends only on the support of S_0, and hence can be interpreted as a worst case result with respect to the signs of S_0. In contrast, our result does not randomize over the signs, but does assume that they are sampled from a fixed sign pattern. Although we do not pursue it here due to space limitations, our analysis also yields a result which holds for worst case sign patterns, and guarantees correct recovery with $\text{rank}(L_0) = O(1)$, and a sparsity pattern of cardinality $\rho n_1 n_2$ for some $\rho > 0$.

that is, we see only those entries $(i, j) \in \Omega_{\text{obs}} \subset [n_1] \times [n_2]$. This models the following problem: we wish to recover L_0 but only see a few entries about L_0, and among those a fraction happens to be corrupted, and we of course do not know which one. As is easily seen, this is a significant extension of the matrix completion problem, which seeks to recover L_0 from undersampled but otherwise perfect data $\mathcal{P}_{\Omega_{\text{obs}}} L_0$.

We propose recovering L_0 by solving the following problem:

$$\begin{aligned} \text{minimize} \quad & \|L\|_* + \lambda \|S\|_1 \\ \text{subject to} \quad & \mathcal{P}_{\Omega_{\text{obs}}}(L + S) = Y. \end{aligned} \tag{1.5}$$

In words, among all decompositions matching the available data, Principal Component Pursuit finds the one that minimizes the weighted combination of the nuclear norm, and of the ℓ_1 norm. Our observation is that under some conditions, this simple approach recovers the low-rank component exactly. In fact, the techniques developed in this paper establish this result:

Theorem 1.2 *Suppose L_0 is $n \times n$, obeys the conditions (1.2)–(1.3), and that Ω_{obs} is uniformly distributed among all sets of cardinality m obeying $m = 0.1n^2$. Suppose for simplicity, that each observed entry is corrupted with probability τ independently of the others. Then there is a numerical constant c such that with probability at least $1 - cn^{-10}$, Principal Component Pursuit (1.5) with $\lambda = 1/\sqrt{0.1n}$ is exact, i.e. $\hat{L} = L_0$, provided that*

$$\text{rank}(L_0) \leq \rho_r \, n\mu^{-1}(\log n)^{-2}, \quad \text{and} \quad \tau \leq \tau_s. \tag{1.6}$$

Above, ρ_r and τ_s are positive numerical constants. For general $n_1 \times n_2$ rectangular matrices, PCP with $\lambda = 1/\sqrt{0.1n_{(1)}}$ succeeds from $m = 0.1n_1 n_2$ corrupted entries with probability at least $1 - cn_{(1)}^{-10}$, provided that $\text{rank}(L_0) \leq \rho_r \, n_{(2)}\mu^{-1}(\log n_{(1)})^{-2}$.

In short, perfect recovery from incomplete and corrupted entries is possible by convex optimization.

On the one hand, this result extends our previous result in the following way. If all the entries are available, i.e. $m = n_1 n_2$, then this is Theorem 1.1. On the other hand, it extends matrix completion results. Indeed, if $\tau = 0$, we have a pure matrix completion problem from about a fraction of the total number of entries, and our theorem guarantees perfect recovery as long as r obeys (1.6), which for large values of r, matches the strongest results available. We remark that the recovery is exact, however, via a different algorithm. To be sure, in matrix completion one typically minimizes the nuclear norm $\|L\|_*$ subject to the constraint $\mathcal{P}_{\Omega_{\text{obs}}} L = \mathcal{P}_{\Omega_{\text{obs}}} L_0$. Here, our program would solve

$$\begin{aligned} \text{minimize} \quad & \|L\|_* + \lambda \|S\|_1 \\ \text{subject to} \quad & \mathcal{P}_{\Omega_{\text{obs}}}(L + S) = \mathcal{P}_{\Omega_{\text{obs}}} L_0, \end{aligned} \tag{1.7}$$

and return $\hat{L} = L_0$, $\hat{S} = 0$! In this context, Theorem 1.2 proves that matrix completion is stable vis a vis gross errors.

Remark. We have stated Theorem 1.2 merely to explain how our ideas can easily be adapted to deal with low-rank matrix recovery problems from undersampled and possibly grossly corrupted data. In our statement, we have chosen to see 10% of the entries but, naturally, similar results hold for all other positive fractions provided that they are large enough. We would like to make it clear that a more careful study is likely to lead to a stronger version of Theorem 1.2. In particular, for very low rank matrices, we expect to see similar results holding with far fewer observations;

that is, in the limit of large matrices, from a decreasing fraction of entries. In fact, our techniques would already establish such sharper results but we prefer not to dwell on such refinements at the moment, and leave this up for future work.

1.7   Notation

We provide a brief summary of the notations used throughout the paper. We shall use five norms of a matrix. The first three are functions of the singular values and they are: 1) the operator norm or 2-norm denoted by $\|X\|$; 2) the Frobenius norm denoted by $\|X\|_F$; and 3) the nuclear norm denoted by $\|X\|_*$. The last two are the ℓ_1 and ℓ_∞ norms of a matrix seen as a long vector, and are denoted by $\|X\|_1$ and $\|X\|_\infty$ respectively. The Euclidean inner product between two matrices is defined by the formula $\langle X, Y \rangle := \mathrm{trace}(X^*Y)$, so that $\|X\|_F^2 = \langle X, X \rangle$.

Further, we will also manipulate linear transformations which act on the space of matrices, and we will use calligraphic letters for these operators as in $\mathcal{P}_\Omega X$. We shall also abuse notation by also letting Ω be the linear space of matrices supported on Ω. Then $\mathcal{P}_{\Omega^\perp}$ denotes the projection onto the space of matrices supported on Ω^c so that $\mathcal{I} = \mathcal{P}_\Omega + \mathcal{P}_{\Omega^\perp}$, where $\mathcal{I}$ is the identity operator. We will consider a single norm for these, namely, the operator norm (the top singular value) denoted by $\|\mathcal{A}\|$, which we may want to think of as $\|\mathcal{A}\| = \sup_{\{\|X\|_F = 1\}} \|\mathcal{A}X\|_F$; for instance, $\|\mathcal{P}_\Omega\| = 1$ whenever $\Omega \neq \emptyset$.

1.8   Organization of the paper

The paper is organized as follows. In Section 2, we provide the key steps in the proof of Theorem 1.1. This proof depends upon on two critical properties of dual certificates, which are established in the separate Section 3. The reason why this is separate is that in a first reading, the reader might want to jump to Section 4, which presents applications to video surveillance, and computer vision. Section 5 introduces algorithmic ideas to find the Principal Component Pursuit solution when M is of very large scale. We conclude the paper with a discussion about future research directions in Section 6. Finally, the proof of Theorem 1.2 is in the Appendix, Section 7, together with those of intermediate results.

2   Architecture of the Proof

This section introduces the key steps underlying the proof of our main result, Theorem 1.1. We will prove the result for square matrices for simplicity, and write $n = n_1 = n_2$. Of course, we shall indicate where the argument needs to be modified to handle the general case. Before we start, it is helpful to review some basic concepts and introduce additional notation that shall be used throughout. For a given scalar x, we denote by $\mathrm{sgn}(x)$ the sign of x, which we take to be zero if $x = 0$. By extension, $\mathrm{sgn}(S)$ is the matrix whose entries are the signs of those of S. We recall that any subgradient of the ℓ_1 norm at S_0 supported on Ω, is of the form

$$\mathrm{sgn}(S_0) + F,$$

where F vanishes on Ω, i.e. $\mathcal{P}_\Omega F = 0$, and obeys $\|F\|_\infty \leq 1$.

We will also manipulate the set of subgradients of the nuclear norm. From now on, we will assume that L_0 of rank r has the singular value decomposition $U\Sigma V^*$, where $U, V \in \mathbb{R}^{n \times r}$ just as

in Section 1.3. Then any subgradient of the nuclear norm at L_0 is of the form

$$UV^* + W,$$

where $U^*W = 0$, $WV = 0$ and $\|W\| \leq 1$. Denote by T the linear space of matrices

$$T := \{UX^* + YV^*, \, X, Y \in \mathbb{R}^{n \times r}\}, \tag{2.1}$$

and by $T^\perp$ its orthogonal complement. It is not hard to see that taken together, $U^*W = 0$ and $WV = 0$ are equivalent to $\mathcal{P}_T W = 0$, where $\mathcal{P}_T$ is the orthogonal projection onto T. Another way to put this is $\mathcal{P}_{T^\perp} W = W$. In passing, note that for any matrix M, $\mathcal{P}_{T^\perp} M = (I - UU^*)M(I - VV^*)$, where we recognize that $I - UU^*$ is the projection onto the orthogonal complement of the linear space spanned by the columns of U and likewise for $(I - VV^*)$. A consequence of this simple observation is that for any matrix M, $\|\mathcal{P}_{T^\perp} M\| \leq \|M\|$, a fact that we will use several times in the sequel. Another consequence is that for any matrix of the form $e_i e_j^*$,

$$\|\mathcal{P}_{T^\perp} e_i e_j^*\|_F^2 = \|(I - UU^*)e_i\|^2 \|(I - VV^*)e_j\|^2 \geq (1 - \mu r/n)^2,$$

where we have assumed $\mu r/n \leq 1$. Since $\|\mathcal{P}_T e_i e_j^*\|_F^2 + \|\mathcal{P}_{T^\perp} e_i e_j^*\|_F^2 = 1$, this gives

$$\|\mathcal{P}_T e_i e_j^*\|_F \leq \sqrt{\frac{2\mu r}{n}}. \tag{2.2}$$

For rectangular matrices, the estimate is $\|\mathcal{P}_T e_i e_j^*\|_F \leq \sqrt{\frac{2\mu r}{\min(n_1, n_2)}}$.

Finally, in the sequel we will write that an event holds with high or large probability whenever it holds with probability at least $1 - O(n^{-10})$ (with $n_{(1)}$ in place of n for rectangular matrices).

2.1 An elimination theorem

We begin with a useful definition and an elementary result we shall use a few times.

Definition 2.1 *We will say that S' is a trimmed version of S if $supp(S') \subset supp(S)$ and $S'_{ij} = S_{ij}$ whenever $S'_{ij} \neq 0$.*

In words, a trimmed version of S is obtained by setting some of the entries of S to zero. Having said this, the following intuitive theorem asserts that if Principal Component Pursuit correctly recovers the low-rank and sparse components of $M_0 = L_0 + S_0$, it also correctly recovers the components of a matrix $M'_0 = L_0 + S'_0$ where S'_0 is a trimmed version of S_0. This is intuitive since the problem is somehow easier as there are fewer things to recover.

Theorem 2.2 *Suppose the solution to (1.1) with input data $M_0 = L_0 + S_0$ is unique and exact, and consider $M'_0 = L_0 + S'_0$, where S'_0 is a trimmed version of S_0. Then the solution to (1.1) with input M'_0 is exact as well.*

Proof Write $S'_0 = \mathcal{P}_{\Omega_0} S_0$ for some $\Omega_0 \subset [n] \times [n]$ and let $(\hat{L}, \hat{S})$ be the solution of (1.1) with input $L_0 + S'_0$. Then

$$\|\hat{L}\|_* + \lambda \|\hat{S}\|_1 \leq \|L_0\|_* + \lambda \|\mathcal{P}_{\Omega_0} S_0\|_1$$

and, therefore,

$$\|\hat{L}\|_* + \lambda \|\hat{S}\|_1 + \lambda \|\mathcal{P}_{\Omega_0^\perp} S_0\|_1 \leq \|L_0\|_* + \lambda \|S_0\|_1.$$

Note that $(\hat{L}, \hat{S} + \mathcal{P}_{\Omega_0^\perp} S_0)$ is feasible for the problem with input data $L_0 + S_0$, and since $\|\hat{S} + \mathcal{P}_{\Omega_0^\perp} S_0\|_1 \le \|\hat{S}\|_1 + \|\mathcal{P}_{\Omega_0^\perp} S_0\|_1$, we have

$$\|\hat{L}\|_* + \lambda\|\hat{S} + \mathcal{P}_{\Omega_0^\perp} S_0\|_1 \le \|L_0\|_* + \lambda\|S_0\|_1.$$

The right-hand side, however, is the optimal value, and by unicity of the optimal solution, we must have $\hat{L} = L_0$, and $\hat{S} + \mathcal{P}_{\Omega_0^\perp} S_0 = S_0$ or $\hat{S} = \mathcal{P}_{\Omega_0} S_0 = S_0'$. This proves the claim. ∎

The Bernoulli model. In Theorem 1.1, probability is taken with respect to the uniformly random subset $\Omega = \{(i,j) : S_{ij} \ne 0\}$ of cardinality m. In practice, it is a little more convenient to work with the *Bernoulli model* $\Omega = \{(i,j) : \delta_{ij} = 1\}$, where the δ_{ij}'s are i.i.d. variables Bernoulli taking value one with probability ρ and zero with probability $1 - \rho$, so that the expected cardinality of Ω is ρn^2. From now on, we will write $\Omega \sim \mathrm{Ber}(\rho)$ as a shorthand for Ω is sampled from the Bernoulli model with parameter ρ.

Since by Theorem 2.2, the success of the algorithm is monotone in $|\Omega|$, any guarantee proved for the Bernoulli model holds for the uniform model as well, and vice versa, if we allow for a vanishing shift in ρ around m/n^2. The arguments underlying this equivalence are standard, see [9, 10], and may be found in the Appendix for completeness.

2.2 Derandomization

In Theorem 1.1, the values of the nonzero entries of S_0 are fixed. It turns out that it is easier to prove the theorem under a stronger assumption, which assumes that the signs of the nonzero entries are independent symmetric Bernoulli variables, i.e. take the value ± 1 with probability $1/2$ (independently of the choice of the support set). The convenient theorem below shows that establishing the result for random signs is sufficient to claim a similar result for fixed signs.

Theorem 2.3 *Suppose L_0 obeys the conditions of Theorem 1.1 and that the locations of the nonzero entries of S_0 follow the Bernoulli model with parameter $2\rho_s$, and the signs of S_0 are i.i.d. ± 1 as above (and independent from the locations). Then if the PCP solution is exact with high probability, then it is also exact with at least the same probability for the model in which the signs are fixed and the locations are sampled from the Bernoulli model with parameter ρ_s.*

This theorem is convenient because to prove our main result, we only need to show that it is true in the case where the signs of the sparse component are random.
Proof Consider the model in which the signs are fixed. In this model, it is convenient to think of S_0 as $\mathcal{P}_\Omega S$, for some fixed matrix S, where Ω is sampled from the Bernoulli model with parameter ρ_s. Therefore, S_0 has independent components distributed as

$$(S_0)_{ij} = \begin{cases} S_{ij}, & \text{w. p. } \rho_s, \\ 0, & \text{w. p. } 1 - \rho_s. \end{cases}$$

Consider now a random sign matrix with i.i.d. entries distributed as

$$E_{ij} = \begin{cases} 1, & \text{w. p. } \rho_s, \\ 0, & \text{w. p. } 1 - 2\rho_s, \\ -1, & \text{w. p. } \rho_s, \end{cases}$$

11

and an "elimination" matrix Δ with entries defined by

$$\Delta_{ij} = \begin{cases} 0, & \text{if } E_{ij}[\text{sgn}(S)]_{ij} = -1, \\ 1, & \text{otherwise.} \end{cases}$$

Note that the entries of Δ are independent since they are functions of independent variables.

Consider now $S_0' = \Delta \circ (|S| \circ E)$, where $\circ$ denotes the Hadamard or componentwise product so that, $[S_0']_{ij} = \Delta_{ij} (|S_{ij}|E_{ij})$. Then we claim that S_0' and S_0 have the same distribution. To see why this is true, it suffices by independence to check that the marginals match. For $S_{ij} \neq 0$, we have

$$\begin{aligned} \mathbb{P}([S_0']_{ij} = S_{ij}) &= \mathbb{P}(\Delta_{ij} = 1 \text{ and } E_{ij} = [\text{sgn}(S)]_{ij}) \\ &= \mathbb{P}(E_{ij}[\text{sgn}(S)]_{ij} \neq -1 \text{ and } E_{ij} = [\text{sgn}(S)]_{ij}) \\ &= \mathbb{P}(E_{ij} = [\text{sgn}(S)]_{ij}) = \rho_s, \end{aligned}$$

which establishes the claim.

This construction allows to prove the theorem. Indeed, $|S| \circ E$ now obeys the random sign model, and by assumption, PCP recovers $|S| \circ E$ with high probability. By the elimination theorem, this program also recovers $S_0' = \Delta \circ (|S| \circ E)$. Since S_0' and S_0 have the same distribution, the theorem follows. ∎

2.3 Dual certificates

We introduce a simple condition for the pair (L_0, S_0) to be the unique optimal solution to Principal Component Pursuit. These conditions are stated in terms of a dual vector, the existence of which certifies optimality. (Recall that Ω is the space of matrices with the same support as the sparse component S_0, and that T is the space defined via the the column and row spaces of the low-rank component L_0 (2.1).)

Lemma 2.4 *Assume that $\|\mathcal{P}_\Omega \mathcal{P}_T\| < 1$. With the standard notations, (L_0, S_0) is the unique solution if there is a pair (W, F) obeying*

$$UV^* + W = \lambda(sgn(S_0) + F),$$

with $\mathcal{P}_T W = 0$, $\|W\| < 1$, $\mathcal{P}_\Omega F = 0$ and $\|F\|_\infty < 1$.

Note that the condition $\|\mathcal{P}_\Omega \mathcal{P}_T\| < 1$ is equivalent to saying that $\Omega \cap T = \{0\}$.

Proof We consider a feasible perturbation $(L_0 + H, S_0 - H)$ and show that the objective increases whenever $H \neq 0$, hence proving that (L_0, S_0) is the unique solution. To do this, let $UV^* + W_0$ be an arbitrary subgradient of the nuclear norm at L_0, and $\text{sgn}(S_0) + F_0$ be an arbitrary subgradient of the ℓ_1-norm at S_0. By definition of subgradients,

$$\|L_0 + H\|_* + \lambda\|S_0 - H\|_1 \geq \|L_0\|_* + \lambda\|S_0\|_1 + \langle UV^* + W_0, H \rangle - \lambda\langle \text{sgn}(S_0) + F_0, H \rangle.$$

Now pick W_0 such that $\langle W_0, H \rangle = \|\mathcal{P}_{T^\perp} H\|_*$ and F_0 such that $\langle F_0, H \rangle = -\|\mathcal{P}_{\Omega^\perp} H\|_1$.[7] We have

$$\|L_0 + H\|_* + \lambda\|S_0 - H\|_1 \geq \|L_0\|_* + \lambda\|S_0\|_1 + \|\mathcal{P}_{T^\perp} H\|_* + \lambda\|\mathcal{P}_{\Omega^\perp} H\|_1 + \langle UV^* - \lambda\text{sgn}(S_0), H \rangle.$$

---

[7]For instance, $F_0 = -\text{sgn}(\mathcal{P}_{\Omega^\perp} H)$ is such a matrix. Also, by duality between the nuclear and the operator norm, there is a matrix obeying $\|W\| = 1$ such that $\langle W, \mathcal{P}_{T^\perp} H \rangle = \|\mathcal{P}_{T^\perp} H\|_*$, and we just take $W_0 = \mathcal{P}_{T^\perp}(W)$.

By assumption

$$|\langle UV^* - \lambda \operatorname{sgn}(S_0), H \rangle| \leq |\langle W, H \rangle| + \lambda |\langle F, H \rangle| \leq \beta (\|\mathcal{P}_{T^\perp} H\|_* + \lambda \|\mathcal{P}_{\Omega^\perp} H\|_1)$$

for $\beta = \max(\|W\|, \|F\|_\infty) < 1$ and, thus,

$$\|L_0 + H\|_* + \lambda \|S_0 - H\|_1 \geq \|L_0\|_* + \lambda \|S_0\|_1 + (1 - \beta)\Big(\|\mathcal{P}_{T^\perp} H\|_* + \lambda \|\mathcal{P}_{\Omega^\perp} H\|_1\Big).$$

Since by assumption, $\Omega \cap T = \{0\}$, we have $\|\mathcal{P}_{T^\perp} H\|_* + \lambda \|\mathcal{P}_{\Omega^\perp} H\|_1 > 0$ unless $H = 0$. ∎

Hence, we see that to prove exact recovery, it is sufficient to produce a 'dual certificate' W obeying

$$\begin{cases} W \in T^\perp, \\ \|W\| < 1, \\ \mathcal{P}_\Omega(UV^* + W) = \lambda \operatorname{sgn}(S_0), \\ \|\mathcal{P}_{\Omega^\perp}(UV^* + W)\|_\infty < \lambda. \end{cases} \tag{2.3}$$

Our method, however, will produce with high probability a slightly different certificate. The idea is to slightly relax the constraint $\mathcal{P}_\Omega(UV^* + W) = \lambda \operatorname{sgn}(S_0)$, a relaxation that has been introduced by David Gross in [22] in a different context. We prove the following lemma.

Lemma 2.5 *Assume $\|\mathcal{P}_\Omega \mathcal{P}_T\| \leq 1/2$ and $\lambda < 1$. Then with the same notation, (L_0, S_0) is the unique solution if there is a pair (W, F) obeying*

$$UV^* + W = \lambda(sgn(S_0) + F + \mathcal{P}_\Omega D)$$

with $\mathcal{P}_T W = 0$ and $\|W\| \leq \frac{1}{2}$, $\mathcal{P}_\Omega F = 0$ and $\|F\|_\infty \leq \frac{1}{2}$, and $\|\mathcal{P}_\Omega D\|_F \leq \frac{1}{4}$.

Proof Following the proof of Lemma 2.4, we have

$$\begin{aligned} \|L_0 + H\|_* + \lambda \|S_0 - H\|_1 &\geq \|L_0\|_* + \lambda \|S_0\|_1 + \frac{1}{2}\Big(\|\mathcal{P}_{T^\perp} H\|_* + \lambda \|\mathcal{P}_{\Omega^\perp} H\|_1\Big) - \lambda \langle \mathcal{P}_\Omega D, H \rangle \\ &\geq \|L_0\|_* + \lambda \|S_0\|_1 + \frac{1}{2}\Big(\|\mathcal{P}_{T^\perp} H\|_* + \lambda \|\mathcal{P}_{\Omega^\perp} H\|_1\Big) - \frac{\lambda}{4}\|\mathcal{P}_\Omega H\|_F. \end{aligned}$$

Observe now that

$$\begin{aligned} \|\mathcal{P}_\Omega H\|_F &\leq \|\mathcal{P}_\Omega \mathcal{P}_T H\|_F + \|\mathcal{P}_\Omega \mathcal{P}_{T^\perp} H\|_F \\ &\leq \frac{1}{2}\|H\|_F + \|\mathcal{P}_{T^\perp} H\|_F \\ &\leq \frac{1}{2}\|\mathcal{P}_\Omega H\|_F + \frac{1}{2}\|\mathcal{P}_{\Omega^\perp} H\|_F + \|\mathcal{P}_{T^\perp} H\|_F \end{aligned}$$

and, therefore,

$$\|\mathcal{P}_\Omega H\|_F \leq \|\mathcal{P}_{\Omega^\perp} H\|_F + 2\|\mathcal{P}_{T^\perp} H\|_F.$$

In conclusion,

$$\|L_0 + H\|_* + \lambda \|S_0 - H\|_1 \geq \|L_0\|_* + \lambda \|S_0\|_1 + \frac{1}{2}\Big((1 - \lambda)\|\mathcal{P}_{T^\perp} H\|_* + \frac{\lambda}{2}\|\mathcal{P}_{\Omega^\perp} H\|_1\Big),$$

and the term between parenthesis is strictly positive when $H \neq 0$. ∎

As a consequence of Lemma 2.5, it now suffices to produce a dual certificate W obeying

$$
\begin{cases}
W \in T^{\perp}, \\
\|W\| < 1/2, \\
\|\mathcal{P}_{\Omega}(UV^* - \lambda\mathrm{sgn}(S_0) + W)\|_F \leq \lambda/4, \\
\|\mathcal{P}_{\Omega^{\perp}}(UV^* + W)\|_{\infty} < \lambda/2.
\end{cases} \tag{2.4}
$$

Further, we would like to note that the existing literature on matrix completion [8] gives good bounds on $\|\mathcal{P}_{\Omega}\mathcal{P}_T\|$, see Theorem 2.6 in Section 2.5.

2.4 Dual certification via the golfing scheme

In the papers [22, 23], Gross introduces a new scheme, termed the golfing scheme, to construct a dual certificate for the matrix completion problem, i.e. the problem of reconstructing a low-rank matrix from a subset of its entries. In this section, we will adapt this clever golfing scheme, with two important modifications, to our separation problem.

Before we introduce our construction, our model assumes that $\Omega \sim \mathrm{Ber}(\rho)$, or equivalently that $\Omega^c \sim \mathrm{Ber}(1-\rho)$. Now the distribution of Ω^c is the same as that of $\Omega^c = \Omega_1 \cup \Omega_2 \cup \ldots \cup \Omega_{j_0}$, where each Ω_j follows the Bernoulli model with parameter q, which has an explicit expression. To see this, observe that by independence, we just need to make sure that any entry (i,j) is selected with the right probability. We have

$$
\mathbb{P}((i,j) \in \Omega) = \mathbb{P}(\mathrm{Bin}(j_0, q) = 0) = (1-q)^{j_0},
$$

so that the two models are the same if

$$
\rho = (1-q)^{j_0},
$$

hence justifying our assertion. Note that because of overlaps between the Ω_j's, $q \geq (1-\rho)/j_0$.

We now propose constructing a dual certificate

$$
W = W^L + W^S,
$$

where each component is as follows:

1. *Construction of W^L via the golfing scheme.* Fix an integer $j_0 \geq 1$ whose value shall be discussed later, and let Ω_j, $1 \leq j \leq j_0$, be defined as above so that $\Omega^c = \cup_{1 \leq j \leq j_0} \Omega_j$. Then starting with $Y_0 = 0$, inductively define

$$
Y_j = Y_{j-1} + q^{-1}\mathcal{P}_{\Omega_j}\mathcal{P}_T(UV^* - Y_{j-1}),
$$

and set

$$
W^L = \mathcal{P}_{T^{\perp}}Y_{j_0}. \tag{2.5}
$$

This is a variation on the golfing scheme discussed in [22], which assumes that the Ω_j's are sampled with replacement, and does not use the projector $\mathcal{P}_{\Omega_j}$ but something more complicated taking into account the number of times a specific entry has been sampled.

14

2. *Construction of W^S via the method of least squares.* Assume that $\|\mathcal{P}_\Omega \mathcal{P}_T\| < 1/2$. Then $\|\mathcal{P}_\Omega \mathcal{P}_T \mathcal{P}_\Omega\| < 1/4$ and, thus, the operator $\mathcal{P}_\Omega - \mathcal{P}_\Omega \mathcal{P}_T \mathcal{P}_\Omega$ mapping Ω onto itself is invertible; we denote its inverse by $(\mathcal{P}_\Omega - \mathcal{P}_\Omega \mathcal{P}_T \mathcal{P}_\Omega)^{-1}$. We then set

$$W^S = \lambda \mathcal{P}_{T^\perp}(\mathcal{P}_\Omega - \mathcal{P}_\Omega \mathcal{P}_T \mathcal{P}_\Omega)^{-1}\mathrm{sgn}(S_0). \tag{2.6}$$

Clearly, an equivalent definition is via the convergent Neumann series

$$W^S = \lambda \mathcal{P}_{T^\perp} \sum_{k \geq 0}(\mathcal{P}_\Omega \mathcal{P}_T \mathcal{P}_\Omega)^k \mathrm{sgn}(S_0). \tag{2.7}$$

Note that $\mathcal{P}_\Omega W^S = \lambda \mathcal{P}_\Omega(I - \mathcal{P}_T)(\mathcal{P}_\Omega - \mathcal{P}_\Omega \mathcal{P}_T \mathcal{P}_\Omega)^{-1}\mathrm{sgn}(S_0) = \lambda\mathrm{sgn}(S_0)$. With this, the construction has a natural interpretation: one can verify that among all matrices $W \in T^\perp$ obeying $\mathcal{P}_\Omega W = \lambda\mathrm{sgn}(S_0)$, W^S is that with minimum Frobenius norm.

Since both W^L and W^S belong to $T^\perp$ and $\mathcal{P}_\Omega W^S = \lambda\mathrm{sgn}(S_0)$, we will establish that $W^L + W^S$ is a valid dual certificate if it obeys

$$\begin{cases} \|W^L + W^S\| < 1/2, \\ \|\mathcal{P}_\Omega(UV^* + W^L)\|_F \leq \lambda/4, \\ \|\mathcal{P}_{\Omega^\perp}(UV^* + W^L + W^S)\|_\infty < \lambda/2. \end{cases} \tag{2.8}$$

2.5 Key lemmas

We now state three lemmas, which taken collectively, establish our main theorem. The first may be found in [8].

Theorem 2.6 *[8, Theorem 4.1] Suppose Ω_0 is sampled from the Bernoulli model with parameter ρ_0. Then with high probability,*

$$\|\mathcal{P}_T - \rho_0^{-1}\mathcal{P}_T \mathcal{P}_{\Omega_0}\mathcal{P}_T\| \leq \epsilon, \tag{2.9}$$

provided that $\rho_0 \geq C_0\,\epsilon^{-2}\frac{\mu r \log n}{n}$ for some numerical constant $C_0 > 0$ (μ is the incoherence parameter). For rectangular matrices, we need $\rho_0 \geq C_0\,\epsilon^{-2}\frac{\mu r \log n_{(1)}}{n_{(2)}}$.

Among other things, this lemma is important because it shows that $\|\mathcal{P}_\Omega \mathcal{P}_T\| \leq 1/2$, provided $|\Omega|$ is not too large. Indeed, if $\Omega \sim \mathrm{Ber}(\rho)$, we have

$$\|\mathcal{P}_T - (1 - \rho)^{-1}\mathcal{P}_T \mathcal{P}_{\Omega^\perp}\mathcal{P}_T\| \leq \epsilon,$$

with the proviso that $1 - \rho \geq C_0\,\epsilon^{-2}\frac{\mu r \log n}{n}$. Note, however, that since $\mathcal{I} = \mathcal{P}_\Omega + \mathcal{P}_{\Omega^\perp}$,

$$\mathcal{P}_T - (1 - \rho)^{-1}\mathcal{P}_T \mathcal{P}_{\Omega^\perp}\mathcal{P}_T = (1 - \rho)^{-1}(\mathcal{P}_T \mathcal{P}_\Omega \mathcal{P}_T - \rho\mathcal{P}_T)$$

and, therefore, by the triangular inequality

$$\|\mathcal{P}_T \mathcal{P}_\Omega \mathcal{P}_T\| \leq \epsilon(1 - \rho) + \rho\|\mathcal{P}_T\| = \rho + \epsilon(1 - \rho).$$

Since $\|\mathcal{P}_\Omega \mathcal{P}_T\|^2 = \|\mathcal{P}_T \mathcal{P}_\Omega \mathcal{P}_T\|$, we have established the following:

15

Corollary 2.7 *Assume that $\Omega \sim Ber(\rho)$, then $\|\mathcal{P}_\Omega \mathcal{P}_T\|^2 \le \rho + \epsilon$, provided that $1 - \rho \ge C_0\, \epsilon^{-2}\, \frac{\mu r \log n}{n}$, where C_0 is as in Theorem 2.6. For rectangular matrices, the modification is as in Theorem 2.6.*

The lemma below is proved is Section 3.

Lemma 2.8 *Assume that $\Omega \sim Ber(\rho)$ with parameter $\rho \le \rho_s$ for some $\rho_s > 0$. Set $j_0 = 2\lceil \log n \rceil$ (use $\log n_{(1)}$ for rectangular matrices). Then under the other assumptions of Theorem 1.1, the matrix W^L (2.5) obeys*

  (a) $\|W^L\| < 1/4$,

  (b) $\|\mathcal{P}_\Omega(UV^* + W^L)\|_F < \lambda/4$,

  (c) $\|\mathcal{P}_{\Omega^\perp}(UV^* + W^L)\|_\infty < \lambda/4$.

Since $\|\mathcal{P}_\Omega \mathcal{P}_T\| < 1$ with large probability, W^S is well defined and the following holds.

Lemma 2.9 *Assume that S_0 is supported on a set Ω sampled as in Lemma 2.8, and that the signs of S_0 are i.i.d. symmetric (and independent of Ω). Then under the other assumptions of Theorem 1.1, the matrix W^S (2.6) obeys*

  (a) $\|W^S\| < 1/4$,

  (b) $\|\mathcal{P}_{\Omega^\perp} W^S\|_\infty < \lambda/4$.

The proof is also in Section 3. Clearly, W^L and W^S obey (2.8), hence certifying that Principal Component Pursuit correctly recovers the low-rank and sparse components with high probability when the signs of S_0 are random. The earlier "derandomization" argument then establishes Theorem 1.1.

3   Proofs of Dual Certification

This section proves the two crucial estimates, namely, Lemma 2.8 and Lemma 2.9.

3.1   Preliminaries

We begin by recording two results which shall be useful in proving Lemma 2.8. While Theorem 2.6 asserts that with large probability,

$$\|Z - \rho_0^{-1} \mathcal{P}_T \mathcal{P}_{\Omega_0} Z\|_F \le \epsilon \|Z\|_F,$$

for all $Z \in T$, the next lemma shows that for a fixed Z, the sup-norm of $Z - \rho_0^{-1} \mathcal{P}_T \mathcal{P}_{\Omega_0}(Z)$ also does not increase (also with large probability).

Lemma 3.1 *Suppose $Z \in T$ is a fixed matrix, and $\Omega_0 \sim Ber(\rho_0)$. Then with high probability,*

$$\|Z - \rho_0^{-1} \mathcal{P}_T \mathcal{P}_{\Omega_0} Z\|_\infty \le \epsilon \|Z\|_\infty \tag{3.1}$$

provided that $\rho_0 \ge C_0\, \epsilon^{-2}\, \frac{\mu r \log n}{n}$ (for rectangular matrices, $\rho_0 \ge C_0\, \epsilon^{-2}\, \frac{\mu r \log n_{(1)}}{n_{(2)}}$) for some numerical constant $C_0 > 0$.

The proof is an application of Bernstein's inequality and may be found in the Appendix. A similar but somewhat different version of (3.1) appears in [44].

The second result was proved in [8].

Lemma 3.2 *[8, Theorem 6.3] Suppose Z is fixed, and $\Omega_0 \sim Ber(\rho_0)$. Then with high probability,*

$$\|(I - \rho_0^{-1}\mathcal{P}_{\Omega_0})Z\| \leq C_0' \sqrt{\frac{n \log n}{\rho_0}} \|Z\|_\infty \tag{3.2}$$

for some small numerical constant $C_0' > 0$ provided that $\rho_0 \geq C_0 \frac{\mu \log n}{n}$ (or $\rho_0 \geq C_0' \frac{\mu \log n_{(1)}}{n_{(2)}}$ for rectangular matrices in which case $n_{(1)} \log n_{(1)}$ replaces $n \log n$ in (3.2)).

As a remark, Lemmas 3.1 and 3.2, and Theorem 2.6 all hold with probability at least $1 - O(n^{-\beta})$, $\beta > 2$, if C_0 is replaced by $C\beta$ for some numerical constant $C > 0$.

3.2 Proof of Lemma 2.8

We begin by introducing a piece of notation and set $Z_j = UV^* - \mathcal{P}_T Y_j$ obeying

$$Z_j = (P_T - q^{-1}\mathcal{P}_T\mathcal{P}_{\Omega_j}\mathcal{P}_T)Z_{j-1}.$$

Obviously $Z_j \in T$ for all $j \geq 0$. First, note that when

$$q \geq C_0 \, \epsilon^{-2} \, \frac{\mu r \log n}{n}, \tag{3.3}$$

(for rectangular matrices, take $q \geq C_0 \epsilon^{-2} \frac{\mu r \log n_{(1)}}{n_{(2)}}$), we have

$$\|Z_j\|_\infty \leq \epsilon\|Z_{j-1}\|_\infty \tag{3.4}$$

by Lemma 3.1. (This holds with high probability because Ω_j and Z_{j-1} are independent, and this is why the golfing scheme is easy to use.) In particular, this gives that with high probability

$$\|Z_j\|_\infty \leq \epsilon^j \|UV^*\|_\infty.$$

When q obeys the same estimate,

$$\|Z_j\|_F \leq \epsilon\|Z_{j-1}\|_F \tag{3.5}$$

by Theorem 2.6. In particular, this gives that with high probability

$$\|Z_j\|_F \leq \epsilon^j \|UV^*\|_F = \epsilon^j \sqrt{r}. \tag{3.6}$$

Below, we will assume $\epsilon \leq e^{-1}$.

Proof of (a). We prove the first part of the lemma and the argument parallels that in [22], see also [44]. From

$$Y_{j_0} = \sum_j q^{-1} \mathcal{P}_{\Omega_j} Z_{j-1},$$

we deduce

$$\begin{aligned}
\|W^L\| = \|\mathcal{P}_{T^\perp} Y_{j_0}\| &\le \sum_j \|q^{-1} \mathcal{P}_{T^\perp} \mathcal{P}_{\Omega_j} Z_{j-1}\| \\
&= \sum_j \|\mathcal{P}_{T^\perp}(q^{-1} \mathcal{P}_{\Omega_j} Z_{j-1} - Z_{j-1})\| \\
&\le \sum_j \|q^{-1} \mathcal{P}_{\Omega_j} Z_{j-1} - Z_{j-1}\| \\
&\le C_0' \sqrt{\frac{n \log n}{q}} \sum_j \|Z_{j-1}\|_\infty \\
&\le C_0' \sqrt{\frac{n \log n}{q}} \sum_j \epsilon^{j-1} \|UV^*\|_\infty \\
&\le C_0'(1-\epsilon)^{-1} \sqrt{\frac{n \log n}{q}} \|UV^*\|_\infty.
\end{aligned}$$

The fourth step follows from Lemma 3.2 and the fifth from (3.5). Since $\|UV^*\|_\infty \le \sqrt{\mu r}/n$, this gives

$$\|W^L\| \le C'\epsilon$$

for some numerical constant C' whenever q obeys (3.3).

Proof of (b). Since $\mathcal{P}_\Omega Y_{j_0} = 0$,

$$\mathcal{P}_\Omega(UV^* + \mathcal{P}_{T^\perp} Y_{j_0}) = \mathcal{P}_\Omega(UV^* - \mathcal{P}_T Y_{j_0}) = \mathcal{P}_\Omega(Z_{j_0}),$$

and it follows from (3.6) that

$$\|Z_{j_0}\|_F \le \epsilon^{j_0} \|UV^*\|_F = \epsilon^{j_0} \sqrt{r}.$$

Since $\epsilon \le e^{-1}$ and $j_0 \ge 2 \log n$, $\epsilon^{j_0} \le 1/n^2$ and this proves the claim.

Proof of (c). We have $UV^* + W^L = Z_{j_0} + Y_{j_0}$ and know that Y_{j_0} is supported on Ω^c. Therefore, since $\|Z_{j_0}\|_F \le \lambda/8$, it suffices to show that $\|Y_{j_0}\|_\infty \le \lambda/8$. We have

$$\begin{aligned}
\|Y_{j_0}\|_\infty &\le q^{-1} \sum_j \|\mathcal{P}_{\Omega_j} Z_{j-1}\|_\infty \\
&\le q^{-1} \sum_j \|Z_{j-1}\|_\infty \\
&\le q^{-1} \sum_j \epsilon^j \|UV^*\|_\infty.
\end{aligned}$$

18

Since $\|UV^*\|_\infty \le \sqrt{\mu r}/n$, this gives

$$\|Y_{j_0}\|_\infty \le C' \frac{\epsilon^2}{\sqrt{\mu r (\log n)^2}}$$

for some numerical constant C' whenever q obeys (3.3). Since $\lambda = 1/\sqrt{n}$, $\|Y_{j_0}\|_\infty \le \lambda/8$ if

$$\epsilon \le C \Big(\frac{\mu r (\log n)^2}{n} \Big)^{1/4}.$$

Summary. We have seen that (a) and (b) are satisfied if ϵ is sufficiently small and $j_0 \ge 2\log n$. For (c), we can take ϵ on the order of $(\mu r (\log n)^2/n)^{1/4}$, which will be sufficiently small as well provided that ρ_r in (1.4) is sufficiently small. Note that everything is consistent since $C_0 \epsilon^{-2} \frac{\mu r \log n}{n} < 1$. This concludes the proof of Lemma 2.8.

3.3 Proof of Lemma 2.9

It is convenient to introduce the sign matrix $E = \mathrm{sgn}(S_0)$ distributed as

$$E_{ij} = \begin{cases} 1, & \text{w. p. } \rho/2, \\ 0, & \text{w. p. } 1 - \rho, \\ -1, & \text{w. p. } \rho/2. \end{cases} \tag{3.7}$$

We shall be interested in the event $\{\|\mathcal{P}_\Omega \mathcal{P}_T\| \le \sigma\}$ which holds with large probability when $\sigma = \sqrt{\rho} + \epsilon$, see Corollary 2.7. In particular, for any $\sigma > 0$, $\{\|\mathcal{P}_\Omega \mathcal{P}_T\| \le \sigma\}$ holds with high probability provided ρ is sufficiently small.

Proof of (a). By construction,

$$W^S = \lambda \mathcal{P}_{T^\perp} E + \lambda \mathcal{P}_{T^\perp} \sum_{k \ge 1} (\mathcal{P}_\Omega \mathcal{P}_T \mathcal{P}_\Omega)^k E$$
$$:= \mathcal{P}_{T^\perp} W_0^S + \mathcal{P}_{T^\perp} W_1^S.$$

For the first term, we have $\|\mathcal{P}_{T^\perp} W_0^S\| \le \|W_0^S\| = \lambda \|E\|$. Then standard arguments about the norm of a matrix with i.i.d. entries give [48]

$$\|E\| \le 4\sqrt{n\rho}$$

with large probability. Since $\lambda = 1/\sqrt{n}$, this gives $\|W_0^S\| \le 4\sqrt{\rho}$. When the matrix is rectangular, we have

$$\|E\| \le 4\sqrt{n_{(1)}\rho}$$

with high probability. Since $\lambda = 1/\sqrt{n_{(1)}}$ in this case, $\|W_0^S\| \le 4\sqrt{\rho}$ as well.

Set $\mathcal{R} = \sum_{k \ge 1} (\mathcal{P}_\Omega \mathcal{P}_T \mathcal{P}_\Omega)^k$ and observe that $\mathcal{R}$ is self-adjoint. For the second term, $\|\mathcal{P}_{T^\perp} W_1^S\| \le \|W_1^S\|$, where $W_1^S = \lambda \mathcal{R}(E)$. We need to bound the operator norm of the matrix $\mathcal{R}(E)$, and use a standard covering argument to do this. Throughout, N denotes an $1/2$-net for $\mathbb{S}^{n-1}$ of size at most 6^n (such a net exists, see [31, Theorem 4.16]). Then a standard argument [48] shows that

$$\|\mathcal{R}(E)\| = \sup_{x,y \in \mathbb{S}^{n-1}} \langle y, \mathcal{R}(E)x \rangle \le 4 \sup_{x,y \in N} \langle y, \mathcal{R}(E)x \rangle.$$

19

For a fixed pair (x, y) of unit-normed vectors in $N \times N$, define the random variable

$$X(x, y) := \langle y, \mathcal{R}(E)x \rangle = \langle \mathcal{R}(yx^*), E \rangle.$$

Conditional on $\Omega = \operatorname{supp}(E)$, the signs of E are i.i.d. symmetric and Hoeffding's inequality gives

$$\mathbb{P}(|X(x, y)| > t \,|\, \Omega) \leq 2 \exp\left(-\frac{2t^2}{\|\mathcal{R}(xy^*)\|_F^2}\right).$$

Now since $\|yx^*\|_F = 1$, the matrix $\mathcal{R}(yx^*)$ obeys $\|\mathcal{R}(yx^*)\|_F \leq \|\mathcal{R}\|$ and, therefore,

$$\mathbb{P}\left(\sup_{x,y \in N} |X(x, y)| > t \,|\, \Omega\right) \leq 2|N|^2 \exp\left(-\frac{2t^2}{\|\mathcal{R}\|^2}\right).$$

Hence,

$$\mathbb{P}(\|\mathcal{R}(E)\| > t \,|\, \Omega) \leq 2|N|^2 \exp\left(-\frac{t^2}{8\|\mathcal{R}\|^2}\right).$$

On the event $\{\|\mathcal{P}_\Omega \mathcal{P}_T\| \leq \sigma\}$,

$$\|\mathcal{R}\| \leq \sum_{k \geq 1} \sigma^{2k} = \frac{\sigma^2}{1 - \sigma^2}$$

and, therefore, unconditionally,

$$\mathbb{P}(\|\mathcal{R}(E)\| > t) \leq 2|N|^2 \exp\left(-\frac{\gamma^2 t^2}{2}\right) + \mathbb{P}(\|\mathcal{P}_\Omega \mathcal{P}_T\| \geq \sigma), \qquad \gamma = \frac{1 - \sigma^2}{2\sigma^2}.$$

This gives

$$\mathbb{P}(\lambda \|\mathcal{R}(E)\| > t) \leq 2 \times 6^{2n} \exp\left(-\frac{\gamma^2 t^2}{2\lambda^2}\right) + \mathbb{P}(\|\mathcal{P}_\Omega \mathcal{P}_T\| \geq \sigma).$$

With $\lambda = 1/\sqrt{n}$,

$$\|W^S\| \leq 1/4,$$

with large probability, provided that σ, or equivalently ρ, is small enough.

Proof of (b). Observe that

$$\mathcal{P}_{\Omega^\perp} W^S = -\lambda \mathcal{P}_{\Omega^\perp} \mathcal{P}_T (\mathcal{P}_\Omega - \mathcal{P}_\Omega \mathcal{P}_T \mathcal{P}_\Omega)^{-1} E.$$

Now for $(i, j) \in \Omega^c$, $W_{ij}^S = \langle e_i, W^S e_j \rangle = \langle e_i e_j^*, W^S \rangle$, and we have

$$W_{ij}^S = \lambda \langle X(i, j), E \rangle,$$

where $X(i, j)$ is the matrix $-(\mathcal{P}_\Omega - \mathcal{P}_\Omega \mathcal{P}_T \mathcal{P}_\Omega)^{-1} \mathcal{P}_\Omega \mathcal{P}_T (e_i e_j^*)$. Conditional on $\Omega = \operatorname{supp}(E)$, the signs of E are i.i.d. symmetric, and Hoeffding's inequality gives

$$\mathbb{P}(|W_{ij}^S| > t\lambda \,|\, \Omega) \leq 2 \exp\left(-\frac{2t^2}{\|X(i, j)\|_F^2}\right),$$

and, thus,

$$\mathbb{P}\left(\sup_{i,j} |W_{ij}^S| > t\lambda \,|\, \Omega\right) \leq 2n^2 \exp\left(-\frac{2t^2}{\sup_{i,j} \|X(i, j)\|_F^2}\right).$$

Since (2.2) holds, we have

$$\|\mathcal{P}_\Omega \mathcal{P}_T(e_i e_j^*)\|_F \leq \|\mathcal{P}_\Omega \mathcal{P}_T\| \|\mathcal{P}_T(e_i e_j^*)\|_F \leq \sigma \sqrt{2\mu r/n}$$

on the event $\{\|\mathcal{P}_\Omega \mathcal{P}_T\| \leq \sigma\}$. On the same event, $\|(\mathcal{P}_\Omega - \mathcal{P}_\Omega \mathcal{P}_T \mathcal{P}_\Omega)^{-1}\| \leq (1 - \sigma^2)^{-1}$ and, therefore,

$$\|X(i,j)\|_F^2 \leq \frac{2\sigma^2}{(1-\sigma^2)^2} \frac{\mu r}{n}.$$

Then unconditionally,

$$\mathbb{P}\Big(\sup_{i,j} |W_{ij}^S| > t\lambda\Big) \leq 2n^2 \exp\Big(-\frac{n\gamma^2 t^2}{\mu r}\Big) + \mathbb{P}(\|\mathcal{P}_\Omega \mathcal{P}_T\| \geq \sigma), \qquad \gamma = \frac{(1-\sigma^2)^2}{2\sigma^2}.$$

This proves the claim when $\mu r < \rho'_r n (\log n)^{-1}$ and ρ'_r is sufficiently small.

4   Numerical Experiments and Applications

In this section, we perform numerical experiments corroborating our main results and suggesting their many applications in image and video analysis. We first investigate Principal Component Pursuit's ability to correctly recover matrices of various rank from errors of various density. We then sketch applications in background modeling from video and removing shadows and specularities from face images.

While the exact recovery guarantee provided by Theorem 1.1 is independent of the particular algorithm used to solve Principal Component Pursuit, its applicability to large scale problems depends on the availability of scalable algorithms for nonsmooth convex optimization. For the experiments in this section, we use the an augmented Lagrange multiplier algorithm introduced in [33, 51].[8] In Section 5, we describe this algorithm in more detail, and explain why it is our algorithm of choice for sparse and low-rank separation.

One important implementation detail in our approach is the choice of λ. Our analysis identifies one choice, $\lambda = 1/\sqrt{\max(n_1, n_2)}$, which works well for incoherent matrices. In order to illustrate the theory, throughout this section we will always choose $\lambda = 1/\sqrt{\max(n_1, n_2)}$. For practical problems, however, it is often possible to improve performance by choosing λ according to prior knowledge about the solution. For example, if we know that S is very sparse, increasing λ will allow us to recover matrices L of larger rank. For practical problems, we recommend $\lambda = 1/\sqrt{\max(n_1, n_2)}$ as a good rule of thumb, which can then be adjusted slightly to obtain the best possible result.

4.1   Exact recovery from varying fractions of error

We first verify the correct recovery phenomenon of Theorem 1.1 on randomly generated problems. We consider square matrices of varying dimension $n = 500, \ldots, 3000$. We generate a rank-r matrix L_0 as a product $L_0 = XY^*$ where X and Y are $n \times r$ matrices with entries independently sampled from a $\mathcal{N}(0, 1/n)$ distribution. S_0 is generated by choosing a support set Ω of size k uniformly at random, and setting $S_0 = \mathcal{P}_\Omega E$, where E is a matrix with independent Bernoulli ± 1 entries.

Table 1 (top) reports the results with $r = \text{rank}(L_0) = 0.05 \times n$ and $k = \|S_0\|_0 = 0.05 \times n^2$. Table 1 (bottom) reports the results for a more challenging scenario, $\text{rank}(L_0) = 0.05 \times n$ and

---

[8]Both [33, 51] have posted a version of their code online.

Dimension n	rank(L_0)	$\|S_0\|_0$	rank($\hat{L}$)	$\|\hat{S}\|_0$	$\frac{\|\hat{L}-L_0\|_F}{\|L_0\|_F}$	# SVD	Time(s)
500	25	12,500	25	12,500	1.1×10^{-6}	16	2.9
1,000	50	50,000	50	50,000	1.2×10^{-6}	16	12.4
2,000	100	200,000	100	200,000	1.2×10^{-6}	16	61.8
3,000	250	450,000	250	450,000	2.3×10^{-6}	15	185.2

$$\text{rank}(L_0) = 0.05 \times n, \ \|S_0\|_0 = 0.05 \times n^2.$$

Dimension n	rank(L_0)	$\|S_0\|_0$	rank($\hat{L}$)	$\|\hat{S}\|_0$	$\frac{\|\hat{L}-L_0\|_F}{\|L_0\|_F}$	# SVD	Time(s)
500	25	25,000	25	25,000	1.2×10^{-6}	17	4.0
1,000	50	100,000	50	100,000	2.4×10^{-6}	16	13.7
2,000	100	400,000	100	400,000	2.4×10^{-6}	16	64.5
3,000	150	900,000	150	900,000	2.5×10^{-6}	16	191.0

$$\text{rank}(L_0) = 0.05 \times n, \ \|S_0\|_0 = 0.10 \times n^2.$$

Table 1: Correct recovery for random problems of varying size. Here, $L_0 = XY^* \in \mathbb{R}^{n \times n}$ with $X, Y \in \mathbb{R}^{n \times r}$; X, Y have entries i.i.d. $\mathcal{N}(0, 1/n)$. $S_0 \in \{-1, 0, 1\}^{n \times n}$ has support chosen uniformly at random and independent random signs; $\|S_0\|_0$ is the number of nonzero entries in S_0. Top: recovering matrices of rank $0.05 \times n$ from 5% gross errors. Bottom: recovering matrices of rank $0.05 \times n$ from 10% gross errors. In all cases, the rank of L_0 and ℓ_0-norm of S_0 are correctly estimated. Moreover, the number of partial singular value decompositions (# SVD) required to solve PCP is almost constant.

$k = 0.10 \times n^2$. In all cases, we set $\lambda = 1/\sqrt{n}$. Notice that in all cases, solving the convex PCP gives a result (L, S) with the correct rank and sparsity. Moreover, the relative error $\|L - L_0\|_F / \|L_0\|_F$ is small, less than 10^{-5} in all examples considered.[9]

The last two columns of Table 1 give the number of partial singular value decompositions computed in the course of the optimization (# SVD) as well as the total computation time. This experiment was performed in Matlab on a Mac Pro with dual quad-core 2.66 GHz Intel Xenon processors and 16 GB RAM. As we will discuss in Section 5 the dominant cost in solving the convex program comes from computing one partial SVD per iteration. Strikingly, in Table 1, the number of SVD computations is nearly constant regardless of dimension, and in all cases less than 17.[10] This suggests that in addition to being theoretically well-founded, the recovery procedure advocated in this paper is also reasonably practical.

4.2 Phase transition in rank and sparsity

Theorem 1.1 shows that convex programming correctly recovers an incoherent low-rank matrix from a constant fraction ρ_s of errors. We next empirically investigate the algorithm's ability to recover matrices of varying rank from errors of varying sparsity. We consider square matrices of

---

[9]We measure relative error in terms of L only, since in this paper we view the sparse and low-rank decomposition as recovering a low-rank matrix L_0 from gross errors. S_0 is of course also well-recovered: in this example, the relative error in S is actually smaller than that in L.

[10]One might reasonably ask whether this near constant number of iterations is due to the fact that random problems are in some sense well-conditioned. There is some validity to this concern, as we will see in our real data examples. [33] suggests a continuation strategy (there termed "Inexact ALM") that produces qualitatively similar solutions with a similarly small number of iterations. However, to the best of our knowledge its convergence is not guaranteed.

dimension $n_1 = n_2 = 400$. We generate low-rank matrices $L_0 = XY^*$ with X and Y independently chosen $n \times r$ matrices with i.i.d. Gaussian entries of mean zero and variance $1/n$. For our first experiment, we assume a Bernoulli model for the support of the sparse term S_0, with random signs: each entry of S_0 takes on value 0 with probability $1 - \rho$, and values ± 1 each with probability $\rho/2$. For each (r, ρ) pair, we generate 10 random problems, each of which is solved via the algorithm of Section 5. We declare a trial to be successful if the recovered $\hat{L}$ satisfies $\|L - L_0\|_F / \|L_0\|_F \leq 10^{-3}$. Figure 1 (left) plots the fraction of correct recoveries for each pair (r, ρ). Notice that there is a large region in which the recovery is exact. This highlights an interesting aspect of our result: the recovery is correct even though in some cases $\|S_0\|_F \gg \|L_0\|_F$ (e.g., for $r/n = \rho$, $\|S_0\|_F$ is $\sqrt{n} = 20$ times larger!). This is to be expected from Lemma 2.4: the existence (or non-existence) of a dual certificate depends only on the signs and support of S_0 and the orientation of the singular spaces of L_0.

However, for incoherent L_0, our main result goes one step further and asserts that the signs of S_0 are also not important: recovery can be guaranteed as long as its support is chosen uniformly at random. We verify this by again sampling L_0 as a product of Gaussian matrices and choosing the support Ω according to the Bernoulli model, but this time setting $S_0 = \mathcal{P}_\Omega \text{sgn}(L_0)$. One might expect such S_0 to be more difficult to distinguish from L_0. Nevertheless, our analysis showed that the number of errors that can be corrected drops by at most $1/2$ when moving to this more difficult model. Figure 1 (middle) plots the fraction of correct recoveries over 10 trials, again varying r and ρ. Interestingly, the region of correct recovery in Figure 1 (middle) actually appears to be broader than that in Figure 1 (left). Admittedly, the shape of the region in the upper-left corner is puzzling, but has been 'confirmed' by several distinct simulation experiments (using different solvers).

Finally, inspired by the connection between matrix completion and robust PCA, we compare the breakdown point for the low-rank and sparse separation problem to the breakdown behavior of the nuclear-norm heuristic for matrix completion. By comparing the two heuristics, we can begin to answer the question *how much is gained by knowing the location Ω of the corrupted entries?* Here, we again generate L_0 as a product of Gaussian matrices. However, we now provide the algorithm with only an incomplete subset $M = \mathcal{P}_{\Omega^\perp} L_0$ of its entries. Each (i, j) is included in Ω independently with probability $1 - \rho$, so rather than a probability of error, here, ρ stands for the probability that an entry is omitted. We solve the nuclear norm minimization problem

$$\text{minimize} \quad \|L\|_* \quad \text{subject to} \quad \mathcal{P}_{\Omega^\perp} L = \mathcal{P}_{\Omega^\perp} M$$

using an augmented Lagrange multiplier algorithm very similar to the one discussed in Section 5. We again declare L_0 to be successfully recovered if $\|L - L_0\|_F / \|L_0\|_F < 10^{-3}$. Figure 1 (right) plots the fraction of correct recoveries for varying r, ρ. Notice that nuclear norm minimization successfully recovers L_0 over a much wider range of (r, ρ). This is interesting because in the regime of large k, $k = \Omega(n^2)$, the best performance guarantees for each heuristic agree in their order of growth – both guarantee correct recovery for $\text{rank}(L_0) = O(n/\log^2 n)$. Fully explaining the difference in performance between the two problems may require a sharper analysis of the breakdown behavior of each.

4.3 Application sketch: background modeling from surveillance video

Video is a natural candidate for low-rank modeling, due to the correlation between frames. One of the most basic algorithmic tasks in video surveillance is to estimate a good model for the
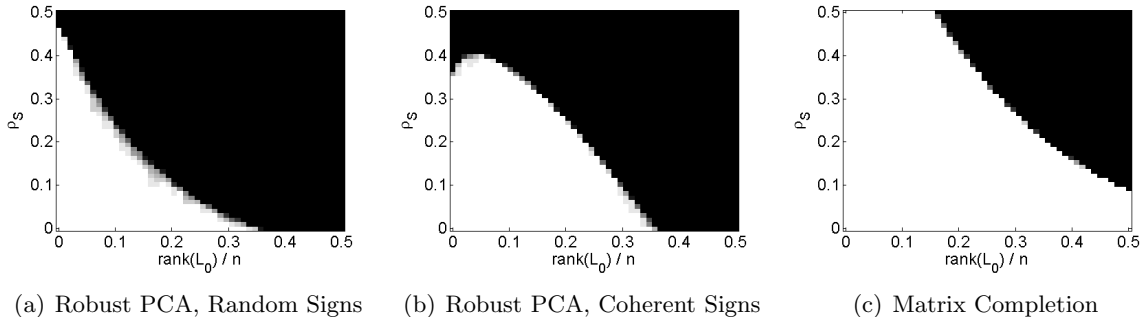
(a) Robust PCA, Random Signs	(b) Robust PCA, Coherent Signs	(c) Matrix Completion

Figure 1: Correct recovery for varying rank and sparsity. Fraction of correct recoveries across 10 trials, as a function of rank(L_0) (x-axis) and sparsity of S_0 (y-axis). Here, $n_1 = n_2 = 400$. In all cases, $L_0 = XY^*$ is a product of independent $n \times r$ i.i.d. $\mathcal{N}(0, 1/n)$ matrices. Trials are considered successful if $\|\hat{L} - L_0\|_F / \|L_0\|_F < 10^{-3}$. Left: low-rank and sparse decomposition, sgn(S_0) random. Middle: low-rank and sparse decomposition, $S_0 = \mathcal{P}_\Omega \text{sgn}(L_0)$. Right: matrix completion. For matrix completion, ρ_s is the probability that an entry is omitted from the observation.

background variations in a scene. This task is complicated by the presence of foreground objects: in busy scenes, every frame may contain some anomaly. Moreover, the background model needs to be flexible enough to accommodate changes in the scene, for example due to varying illumination. In such situations, it is natural to model the background variations as approximately low rank. Foreground objects, such as cars or pedestrians, generally occupy only a fraction of the image pixels and hence can be treated as sparse errors.

We investigate whether convex optimization can separate these sparse errors from the low-rank background. Here, it is important to note that the error support may not be well-modeled as Bernoulli: errors tend to be spatially coherent, and more complicated models such as Markov random fields may be more appropriate [11,52]. Hence, our theorems do not necessarily guarantee the algorithm will succeed with high probability. Nevertheless, as we will see, Principal Component Pursuit still gives visually appealing solutions to this practical low-rank and sparse separation problem, without using any additional information about the spatial structure of the error.

We consider two example videos introduced in [32]. The first is a sequence of 200 grayscale frames taken in an airport. This video has a relatively static background, but significant foreground variations. The frames have resolution 176×144; we stack each frame as a column of our matrix $M \in \mathbb{R}^{25,344 \times 200}$. We decompose M into a low-rank term and a sparse term by solving the convex PCP problem (1.1) with $\lambda = 1/\sqrt{n_1}$. On a desktop PC with a 2.33 GHz Core2 Duo processor and 2 GB RAM, our Matlab implementation requires 806 iterations, and roughly 43 minutes to converge.[11] Figure 2(a) shows three frames from the video; (b) and (c) show the corresponding columns of the low rank matrix $\hat{L}$ and sparse matrix $\hat{S}$ (its absolute value is shown here). Notice that $\hat{L}$ correctly recovers the background, while $\hat{S}$ correctly identifies the moving pedestrians. The person appearing in the images in $\hat{L}$ does not move throughout the video.

Figure 2 (d) and (e) compares the result obtained by Principal Component Pursuit to a state-of-

---

[11]The paper [33] suggests a variant of ALM optimization procedure, there termed the "Inexact ALM" that finds a visually similar decomposition in far fewer iterations (less than 50). However, since the convergence guarantee for that variant is weak, we choose to present the slower, exact result here.
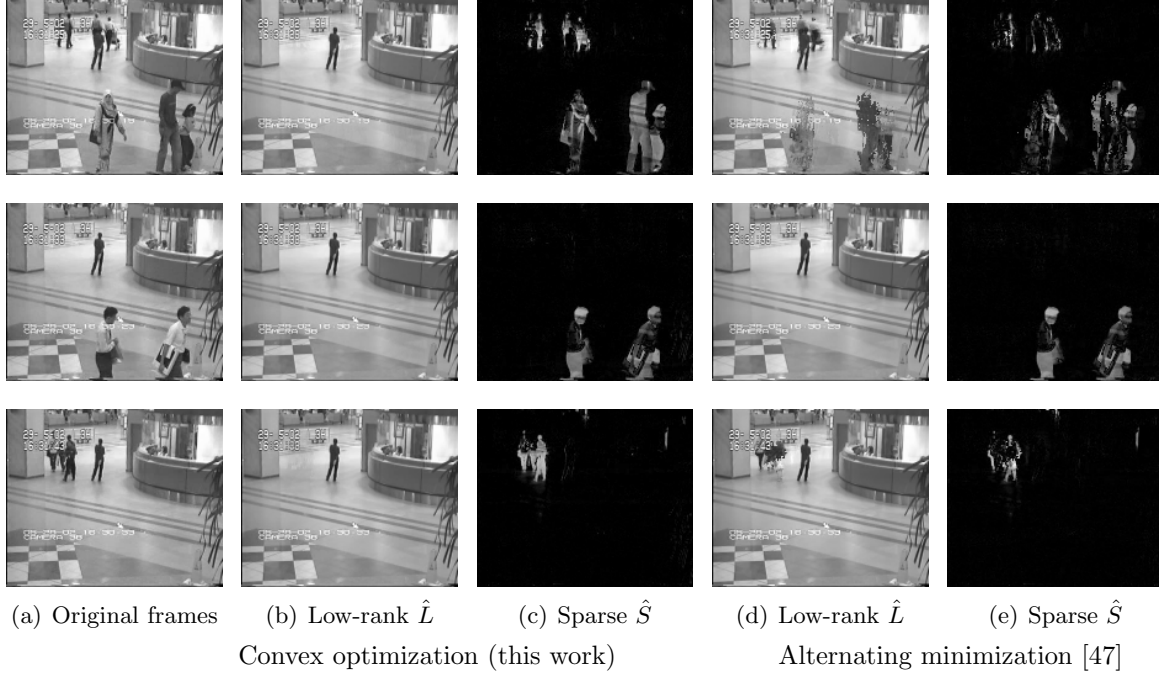
(a) Original frames	(b) Low-rank $\hat{L}$	(c) Sparse $\hat{S}$	(d) Low-rank $\hat{L}$	(e) Sparse $\hat{S}$
	Convex optimization (this work)		Alternating minimization [47]	

Figure 2: Background modeling from video. Three frames from a 200 frame video sequence taken in an airport [32]. (a) Frames of original video M. (b)-(c) Low-rank $\hat{L}$ and sparse components $\hat{S}$ obtained by PCP, (d)-(e) competing approach based on alternating minimization of an m-estimator [47]. PCP yields a much more appealing result despite using less prior knowledge.

the-art technique from the computer vision literature, [47].[12] That approach also aims at robustly recovering a good low-rank approximation, but uses a more complicated, nonconvex m-estimator, which incorporates a local scale estimate that implicitly exploits the spatial characteristics of natural images. This leads to a highly nonconvex optimization, which is solved locally via alternating minimization. Interestingly, despite using more prior information about the signal to be recovered, this approach does not perform as well as the convex programming heuristic: notice the large artifacts in the top and bottom rows of Figure 2 (d).

In Figure 3, we consider 250 frames of a sequence with several drastic illumination changes. Here, the resolution is 168×120, and so M is a $20,160 \times 250$ matrix. For simplicity, and to illustrate the theoretical results obtained above, we again choose $\lambda = 1/\sqrt{n_1}$.[13] For this example, on the same 2.66 GHz Core 2 Duo machine, the algorithm requires a total of 561 iterations and 36 minutes to converge.

Figure 3 (a) shows three frames taken from the original video, while (b) and (c) show the recovered low-rank and sparse components, respectively. Notice that the low-rank component correctly identifies the main illuminations as background, while the sparse part corresponds to the motion in the scene. On the other hand, the result produced by the algorithm of [47] treats some

---

[12] We use the code package downloaded from `http://www.salleurl.edu/~ftorre/papers/rpca/rpca.zip`, modified to choose the rank of the approximation as suggested in [47].

[13] For this example, slightly more appealing results can actually be obtained by choosing larger λ (say, $2/\sqrt{n_1}$).

(a) Original frames   (b) Low-rank $\hat{L}$ (c) Sparse $\hat{S}$ (d) Low-rank $\hat{L}$ (e) Sparse $\hat{S}$

Convex optimization (this work)   Alternating minimization [47]
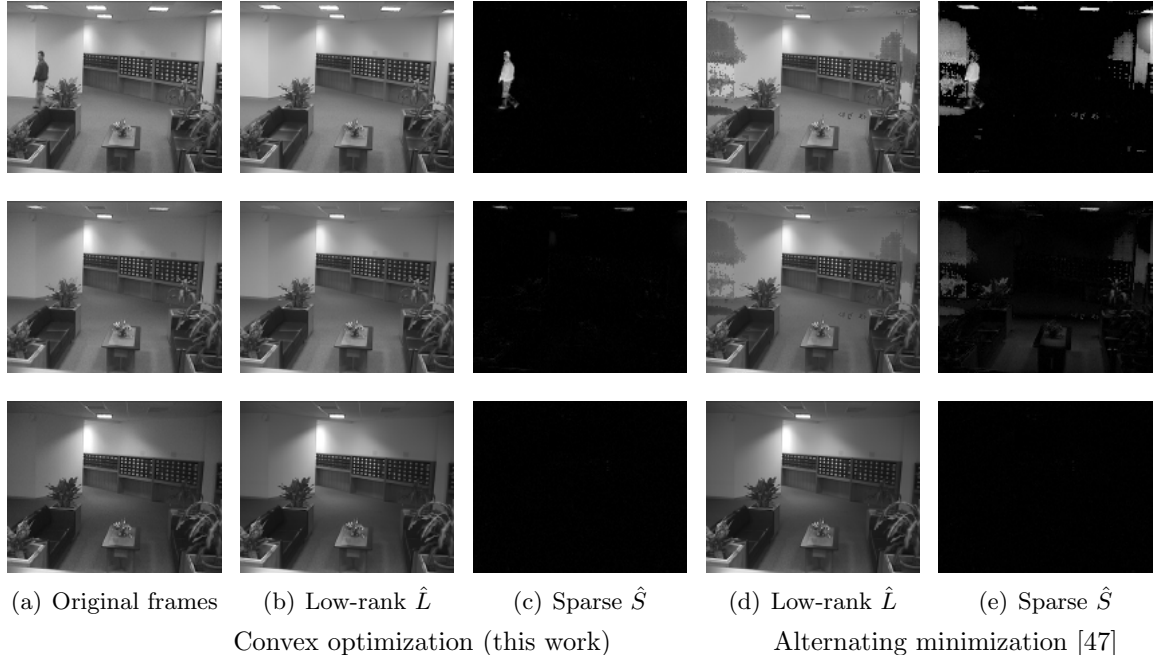
Figure 3: Background modeling from video. Three frames from a 250 frame sequence taken in a lobby, with varying illumination [32]. (a) Original video M. (b)-(c) Low-rank $\hat{L}$ and sparse $\hat{S}$ obtained by PCP. (d)-(e) Low-rank and sparse components obtained by a competing approach based on alternating minimization of an m-estimator [47]. Again, convex programming yields a more appealing result despite using less prior information.

of the first illumination as foreground. PCP again outperforms the competing approach, despite using less prior information. These results suggest the potential power for convex programming as a tool for video analysis.

Notice that the number of iterations for the real data is typically higher than that of the simulations with random matrices given in Table 1. The reason for this discrepancy might be that the structures of real data could slightly deviate from the idealistic low-rank and sparse model. Nevertheless, it is important to realize that practical applications such as video surveillance often provide additional information about the signals of interest, e.g. the support of the sparse foreground is spatially piecewise contiguous, or even impose additional requirements, e.g. the recovered background needs to be non-negative etc. We note that the simplicity of our objective and solution suggests that one can easily incorporate additional constraints and more accurate models of the signals so as to obtain much more efficient and accurate solutions in the future.

4.4 Application sketch: removing shadows and specularities from face images

Face recognition is another problem domain in computer vision where low-dimensional linear models have received a great deal of attention. This is mostly due to the work of Basri and Jacobs, who showed that for convex, Lambertian objects, images taken under distant illumination lie near an approximately nine-dimensional linear subspace known as the *harmonic plane* [1]. However, since faces are neither perfectly convex nor Lambertian, real face images often violate this low-rank
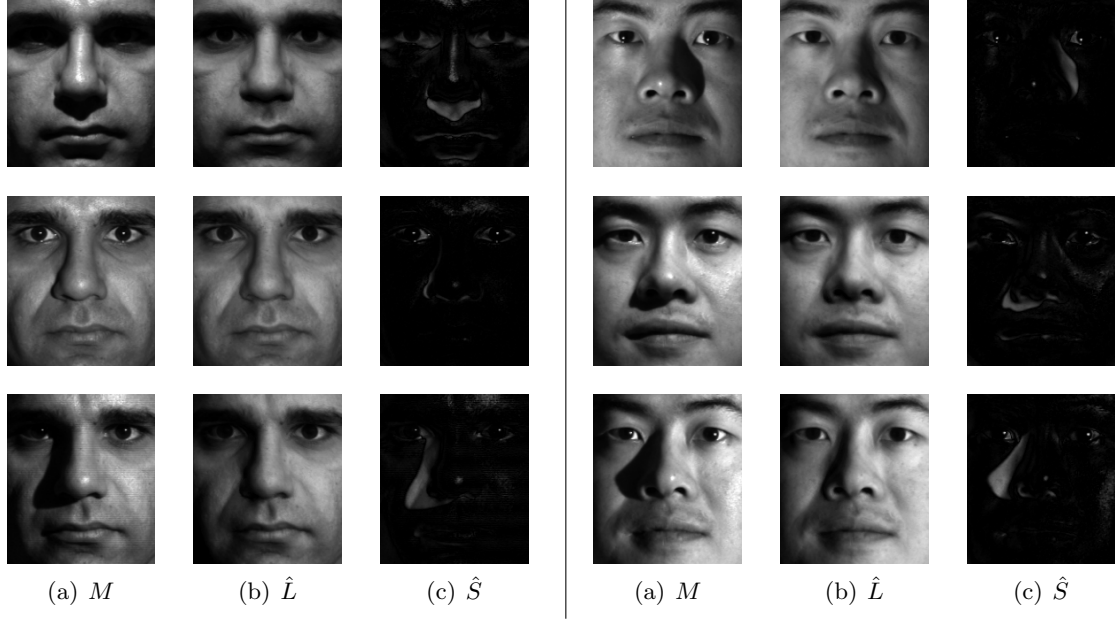
| (a) M | (b) $\hat{L}$ | (c) $\hat{S}$ | | (a) M | (b) $\hat{L}$ | (c) $\hat{S}$ |

Figure 4: Removing shadows, specularities, and saturations from face images. (a) Cropped and aligned images of a person's face under different illuminations from the Extended Yale B database. The size of each image is 192×168 pixels, a total of 58 different illuminations were used for each person. (b) Low-rank approximation $\hat{L}$ recovered by convex programming. (c) Sparse error $\hat{S}$ corresponding to specularities in the eyes, shadows around the nose region, or brightness saturations on the face. Notice in the bottom left that the sparse term also compensates for errors in image acquisition.

model, due to cast shadows and specularities. These errors are large in magnitude, but sparse in the spatial domain. It is reasonable to believe that if we have enough images of the same face, Principal Component Pursuit will be able to remove these errors. As with the previous example, some caveats apply: the theoretical result suggests the performance should be good, but does not guarantee it, since again the error support does not follow a Bernoulli model. Nevertheless, as we will see, the results are visually striking.

Figure 4 shows two examples with face images taken from the Yale B face database [18]. Here, each image has resolution 192×168; there are a total of 58 illuminations per subject, which we stack as the columns of our matrix $M \in \mathbb{R}^{32,256 \times 58}$. We again solve PCP with $\lambda = 1/\sqrt{n_1}$. In this case, the algorithm requires 642 iterations to converge, and the total computation time on the same Core 2 Duo machine is 685 seconds.

Figure 4 plots the low rank term $\hat{L}$ and the magnitude of the sparse term $\hat{S}$ obtained as the solution to the convex program. The sparse term $\hat{S}$ compensates for cast shadows and specular regions. In one example (bottom row of Figure 4 left), this term also compensates for errors in image acquisition. These results may be useful for conditioning the training data for face recognition, as well as face alignment and tracking under illumination variations.

5   Algorithms

Theorem 1.1 shows that incoherent low-rank matrices can be recovered from nonvanishing fractions of gross errors in polynomial time. Moreover, as the experiments in the previous section attest, the low computation cost is guaranteed not only in theory, the efficiency is becoming *practical* for real imaging problems. This practicality is mainly due to the rapid recent progress in scalable algorithms for nonsmooth convex optimization, in particular for minimizing the ℓ_1 and nuclear norms. In this section, we briefly review this progress, and discuss our algorithm of choice for this problem.

For small problem sizes, Principal Component Pursuit

$$
\begin{aligned}
&\text{minimize} &&\|L\|_* + \lambda\|S\|_1 \\
&\text{subject to} &&L + S = M
\end{aligned}
$$

can be performed using off-the-shelf tools such as interior point methods [21]. This was suggested for rank minimization in [16, 45] and for low-rank and sparse decomposition [12] (see also [36]). However, despite their superior convergence rates, interior point methods are typically limited to small problems, say $n < 100$, due to the $O(n^6)$ complexity of computing a step direction.

The limited scalability of interior point methods has inspired a recent flurry of work on first-order methods. Exploiting an analogy with iterative thresholding algorithms for ℓ_1-minimization [49, 50], Cai et. al. developed an algorithm that performs nuclear-norm minimization by repeatedly shrinking the *singular values* of an appropriate matrix, essentially reducing the complexity of each iteration to the cost of an SVD [6]. However, for our low-rank and sparse decomposition problem, this form of iterative thresholding converges slowly, requiring up to 10^4 iterations. Ma et. al. [20, 37] suggest improving convergence using continuation techniques, and also demonstrate how Bregman iterations [42] can be applied to nuclear norm minimization.

The convergence of iterative thresholding has also been greatly improved using ideas from Nesterov's optimal first-order algorithm for smooth minimization [38], which was extended to nonsmooth optimization in [2, 39], and applied to ℓ_1-minimization in [2, 3, 40]. Based on [2], Toh et. al. developed a proximal gradient algorithm for matrix completion which they termed *Accelerated Proximal Gradient (APG)*. A very similar APG algorithm was suggested for low-rank and sparse decomposition in [34]. That algorithm inherits the optimal $O(1/k^2)$ convergence rate for this class of problems. Empirical evidence suggests that these algorithms can solve the convex PCP problem at least 50 times faster than straightforward iterative thresholding (for more details and comparisons, see [34]).

However, despite its good convergence guarantees, the practical performance of APG depends strongly on the design of good continuation schemes. Generic continuation does not guarantee good accuracy and convergence across a wide range of problem settings.[14] In this paper, we have chosen to instead solve the convex PCP problem (1.1) using an augmented Lagrange multiplier (ALM) algorithm introduced in [33, 51]. In our experience, ALM achieves much higher accuracy than APG, in fewer iterations. It works stably across a wide range of problem settings with no tuning of parameters. Moreover we observe an appealing (empirical) property: the rank of the iterates often remains bounded by $\text{rank}(L_0)$ throughout the optimization, allowing them to be computed especially efficiently. APG, on the other hand, does not have this property.

---

[14]In our experience, the optimal choice may depend on the relative magnitudes of the L and S terms and the sparsity of the corruption.

The ALM method operates on the *augmented Lagrangian*

$$l(L, S, Y) = \|L\|_* + \lambda\|S\|_1 + \langle Y, M - L - S \rangle + \frac{\mu}{2}\|M - L - S\|_F^2. \tag{5.1}$$

A generic Lagrange multiplier algorithm [5] would solve PCP by repeatedly setting $(L_k, S_k) = \arg\min_{L,S} l(L, S, Y_k)$, and then updating the Lagrange multiplier matrix via $Y_{k+1} = Y_k + \mu(M - L_k - S_k)$.

For our low-rank and sparse decomposition problem, we can avoid having to solve a sequence of convex programs by recognizing that $\min_L l(L, S, Y)$ and $\min_S l(L, S, Y)$ both have very simple and efficient solutions. Let $\mathcal{S}_\tau : \mathbb{R} \to \mathbb{R}$ denote the shrinkage operator $\mathcal{S}_\tau[x] = \mathrm{sgn}(x)\max(|x| - \tau, 0)$, and extend it to matrices by applying it to each element. It is easy to show that

$$\arg\min_S l(L, S, Y) = \mathcal{S}_{\lambda\mu^{-1}}(M - L + \mu^{-1}Y). \tag{5.2}$$

Similarly, for matrices X, let $\mathcal{D}_\tau(X)$ denote the singular value thresholding operator given by $\mathcal{D}_\tau(X) = U\mathcal{S}_\tau(\Sigma)V^*$, where $X = U\Sigma V^*$ is any singular value decomposition. It is not difficult to show that

$$\arg\min_L l(L, S, Y) = \mathcal{D}_{\mu^{-1}}(M - S - \mu^{-1}Y). \tag{5.3}$$

Thus, a more practical strategy is to first minimize l with respect to L (fixing S), then minimize l with respect to S (fixing L), and then finally update the Lagrange multiplier matrix Y based on the residual $M - L - S$, a strategy that is summarized as Algorithm 1 below.

Algorithm 1 (Principal Component Pursuit by Alternating Directions [33, 51])

1: **initialize:** $S_0 = Y_0 = 0, \mu > 0$.
2: **while** not converged **do**
3:    compute $L_{k+1} = \mathcal{D}_{\mu^{-1}}(M - S_k - \mu^{-1}Y_k)$;
4:    compute $S_{k+1} = \mathcal{S}_{\lambda\mu^{-1}}(M - L_{k+1} + \mu^{-1}Y_k)$;
5:    compute $Y_{k+1} = Y_k + \mu(M - L_{k+1} - S_{k+1})$;
6: **end while**
7: **output:** L, S.

Algorithm 1 is a special case of a more general class of augmented Lagrange multiplier algorithms known as *alternating directions* methods [51]. The convergence of these algorithms has been well-studied (see e.g. [30,35] and the many references therein, as well as discussion in [33,51]). Algorithm 1 performs excellently on a wide range of problems: as we saw in Section 3, relatively small numbers of iterations suffice to achieve good relative accuracy. The dominant cost of each iteration is computing L_{k+1} via singular value thresholding. This requires us to compute those singular vectors of $M - S_k - \mu^{-1}Y_k$ whose corresponding singular values exceed the threshold μ^{-1}. Empirically, we have observed that the number of such large singular values is often bounded by $\mathrm{rank}(L_0)$, allowing the next iterate to be computed efficiently via a partial SVD.[15] The most important implementation details for this algorithm are the choice of μ and the stopping criterion. In this work, we simply choose $\mu = n_1 n_2/4\|M\|_1$, as suggested in [51]. We terminate the algorithm when $\|M - L - S\|_F \leq \delta\|M\|_F$, with $\delta = 10^{-7}$.

---

[15]Further performance gains might be possible by replacing this partial SVD with an approximate SVD, as suggested in [20] for nuclear norm minimization.

Very similar ideas can be used to develop simple and effective augmented Lagrange multiplier algorithms for matrix completion [33], and for the robust matrix completion problem (1.5) discussed in Section 1.6, with similarly good performance. In the preceding section, all simulations and experiments are therefore conducted using ALM-based algorithms. For a more thorough discussion, implementation details and comparisons with other algorithms, please see [33, 51].

6 Discussion

This paper delivers some rather surprising news: one can disentangle the low-rank and sparse components exactly by convex programming, and this provably works under very broad conditions that are much broader than those provided by the best known results. Further, our analysis has revealed rather close relationships between matrix completion and matrix recovery (from sparse errors) and our results even generalize to the case when there are both incomplete and corrupted entries (i.e. Theorem 1.2). In addition, Principal Component Pursuit does not have any free parameter and can be solved by simple optimization algorithms with remarkable efficiency and accuracy. More importantly, our results may point to a very wide spectrum of new theoretical and algorithmic issues together with new practical applications that can now be studied systematically.

Our study so far is limited to the low-rank component being exactly low-rank, and the sparse component being exactly sparse. It would be interesting to investigate when either or both these assumptions are relaxed. One way to think of this is via the new observation model $M = L_0 + S_0 + N_0$, where N_0 is a dense, small perturbation accounting for the fact that the low-rank component is only approximately low-rank and that small errors can be added to all the entries (in some sense, this model unifies the classical PCA and the robust PCA by combining both sparse gross errors and dense small noise). The ideas developed in [7] in connection with the stability of matrix completion under small perturbations may be useful here. Even more generally, the problems of sparse signal recovery, low-rank matrix completion, classical PCA, and robust PCA can all be considered as special cases of a general measurement model of the form

$$M = \mathcal{A}(L_0) + \mathcal{B}(S_0) + \mathcal{C}(N_0),$$

where $\mathcal{A}, \mathcal{B}, \mathcal{C}$ are known linear maps. An ambitious goal might be to understand exactly under what conditions, one can effectively retrieve or decompose L_0 and S_0 from such noisy linear measurements via convex programming.

The remarkable ability of convex optimizations in recovering low-rank matrices and sparse signals in high-dimensional spaces suggest that they will be a powerful tool for processing massive data sets that arise in image/video processing, web data analysis, and bioinformatics. Such data are often of millions or even billions of dimensions so the computational and memory cost can be far beyond that of a typical PC. Thus, one important direction for future investigation is to develop algorithms that have even better scalability, and can be easily implemented on the emerging parallel and distributed computing infrastructures.

7 Appendix

7.1 Equivalence of sampling models

We begin by arguing that a recovery result under the Bernoulli model automatically implies a corresponding result for the uniform model. Denote by $\mathbb{P}_{\text{Unif}(m)}$ and $\mathbb{P}_{\text{Ber}(p)}$ probabilities calculated under the uniform and Bernoulli models and let "Success" be the event that the algorithm succeeds. We have

$$\mathbb{P}_{\text{Ber}(p)}(\text{Success}) = \sum_{k=0}^{n^2} \mathbb{P}_{\text{Ber}(p)}(\text{Success} \mid |\Omega| = k)\, \mathbb{P}_{\text{Ber}(p)}(|\Omega| = k)$$

$$\leq \sum_{k=0}^{m-1} \mathbb{P}_{\text{Ber}(p)}(|\Omega| = k) + \sum_{k=m}^{n^2} \mathbb{P}_{\text{Unif}(k)}(\text{Success})\, \mathbb{P}_{\text{Ber}(p)}(|\Omega| = k)$$

$$\leq \mathbb{P}_{\text{Ber}(p)}(|\Omega| < m) + \mathbb{P}_{\text{Unif}(m)}(\text{Success}),$$

where we have used the fact that for $k \geq m$, $\mathbb{P}_{\text{Unif}(k)}(\text{Success}) \leq \mathbb{P}_{\text{Unif}(m)}(\text{Success})$, and that the conditional distribution of Ω given its cardinality is uniform. Thus,

$$\mathbb{P}_{\text{Unif}(m)}(\text{Success}) \geq \mathbb{P}_{\text{Ber}(p)}(\text{Success}) - \mathbb{P}_{\text{Ber}(p)}(|\Omega| < m).$$

Take $p = m/n^2 + \epsilon$, where $\epsilon > 0$. The conclusion follows from $\mathbb{P}_{\text{Ber}(p)}(|\Omega| < m) \leq e^{-\frac{\epsilon^2 n^2}{2p}}$. In the other direction, the same reasoning gives

$$\mathbb{P}_{\text{Ber}(p)}(\text{Success}) \geq \sum_{k=0}^{m} \mathbb{P}_{\text{Ber}(p)}(\text{Success} \mid |\Omega| = k)\, \mathbb{P}_{\text{Ber}(p)}(|\Omega| = k)$$

$$\geq \mathbb{P}_{\text{Unif}(m)}(\text{Success}) \sum_{k=0}^{m} \mathbb{P}_{\text{Ber}(p)}(|\Omega| = k)$$

$$= \mathbb{P}_{\text{Unif}(m)}(\text{Success})\, \mathbb{P}(|\Omega| \leq m),$$

and choosing m such that $\mathbb{P}(|\Omega| > m)$ is exponentially small, establishes the claim.

7.2 Proof of Lemma 3.1

The proof is essentially an application of Bernstein's inequality, which states that for a sum of uniformly bounded independent random variables with $|Y_k - \mathbb{E}\, Y_k| < c$,

$$\mathbb{P}\Big(\sum_{k=1}^{n}(Y_k - \mathbb{E}\, Y_k) > t\Big) \leq 2\exp\Big(-\frac{t^2}{2\sigma^2 + 2ct/3}\Big), \tag{7.1}$$

where σ^2 is the sum of the variances, $\sigma^2 \equiv \sum_{k=1}^{n} \text{Var}(Y_k)$.

Define Ω_0 via $\Omega_0 = \{(i,j) : \delta_{ij} = 1\}$ where $\{\delta_{ij}\}$ is an independent sequence of Bernoulli variables with parameter ρ_0. With this notation, $Z' = Z - \rho_0^{-1}\mathcal{P}_T\mathcal{P}_{\Omega_0}Z$ is given by

$$Z' = \sum_{ij}(1 - \rho_0^{-1}\delta_{ij})Z_{ij}\mathcal{P}_T(e_i e_j^*)$$

31

so that $Z'_{i_0 j_0}$ is a sum of independent random variables,

$$Z'_{i_0 j_0} = \sum_{ij} Y_{ij}, \qquad Y_{ij} = (1 - \rho_0^{-1} \delta_{ij}) Z_{ij} \langle \mathcal{P}_T(e_i e_j^*), e_{i_0} e_{j_0}^* \rangle.$$

We have

$$\sum_{ij} \operatorname{Var}(Y_{ij}) = (1 - \rho_0) \rho_0^{-1} \sum_{ij} |Z_{ij}|^2 |\langle \mathcal{P}_T(e_i e_j^*), e_{i_0} e_{j_0}^* \rangle|^2$$

$$\leq (1 - \rho_0) \rho_0^{-1} \|Z\|_\infty^2 \sum_{ij} |\langle e_i e_j^*, \mathcal{P}_T(e_{i_0} e_{j_0}^*) \rangle|^2$$

$$= (1 - \rho_0) \rho_0^{-1} \|Z\|_\infty^2 \|\mathcal{P}_T(e_{i_0} e_{j_0}^*)\|_F^2$$

$$\leq (1 - \rho_0) \rho_0^{-1} \|Z\|_\infty^2 \frac{2\mu r}{n},$$

where the last inequality holds because of (2.2). Also, it follows from (1.2) that $|\langle \mathcal{P}_T(e_i e_j^*), e_{i_0} e_{j_0}^* \rangle| \leq \|\mathcal{P}_T(e_i e_j^*)\|_F \|\mathcal{P}_T(e_{i_0} e_{j_0}^*)\|_F \leq 2\mu r/n$ so that $|Y_{ij}| \leq \rho_0^{-1} \|Z\|_\infty \mu r/n$. Then Bernstein's inequality gives

$$\mathbb{P}(|Z'_{ij}| > \epsilon \|Z\|_\infty) \leq 2 \exp\left(-\frac{3}{16} \frac{\epsilon^2 n \rho_0}{\mu r}\right).$$

If ρ_0 is as in Lemma 3.1, the union bound proves the claim.

7.3   Proof of Theorem 1.2

This section presents a proof of Theorem 1.2, which resembles that of Theorem 1.1. Here and below, $S'_0 = \mathcal{P}_{\Omega_{\text{obs}}} S_0$ so that the available data are of the form $Y = \mathcal{P}_{\Omega_{\text{obs}}} L_0 + S'_0$. We make three observations.

- If PCP correctly recovers L_0 from the input data $\mathcal{P}_{\Omega_{\text{obs}}} L_0 + S'_0$ (note that this means that $\hat{L} = L_0$ and $\hat{S} = S'_0$), then it must correctly recover L_0 from $\mathcal{P}_{\Omega_{\text{obs}}} L_0 + S''_0$, where S''_0 is a trimmed version of S'_0. The proof is identical to that of our elimination result, namely, Theorem 2.2. The derandomization argument then applies and it suffices to consider the case where the signs of S'_0 are i.i.d. symmetric Bernoulli variables.

- It is of course sufficient to prove the theorem when each entry in Ω_{obs} is revealed with probability $p_0 := 0.1$, i.e. when $\Omega_{\text{obs}} \sim \operatorname{Ber}(p_0)$.

- We establish the theorem in the case where $n_1 = n_2 = n$ as slight modifications would give the general case.

Further, there are now three index sets of interest:

- Ω_{obs} are those locations where data are available.

- $\Gamma \subset \Omega_{\text{obs}}$ are those locations where data are available and clean; that is, $\mathcal{P}_\Gamma Y = \mathcal{P}_\Gamma L_0$.

- $\Omega = \Omega_{\text{obs}} \setminus \Gamma$ are those locations where data are available but totally unreliable.

The matrix S'_0 is thus supported on Ω. If $\Omega_{\text{obs}} \sim \operatorname{Ber}(p_0)$, then by definition, $\Omega \sim \operatorname{Ber}(p_0 \tau)$.

Dual certification. We begin with two lemmas concerning dual certification.

Lemma 7.1 *Assume $\|\mathcal{P}_{\Gamma^\perp}\mathcal{P}_T\| < 1$. Then (L_0, S_0') is the unique solution if there is a pair (W, F) obeying*

$$UV^* + W = \lambda(sgn(S_0') + F),$$

with $\mathcal{P}_T W = 0$, $\|W\| < 1$, $\mathcal{P}_{\Gamma^\perp} F = 0$ and $\|F\|_\infty < 1$.

The proof is about the same as that of Lemma 2.4, and is discussed in very brief terms. The idea is to consider a feasible perturbation of the form $(L_0 + H_L, S_0' - H_S)$ obeying $\mathcal{P}_{\Omega_{\mathrm{obs}}} H_L = \mathcal{P}_{\Omega_{\mathrm{obs}}} H_S$, and show that this increases the objective functional unless $H_L = H_S = 0$. Then a sequence of steps similar to that in the proof of Lemma 2.4 establishes

$$\|L_0 + H_L\|_* + \lambda\|S_0' - H_S\|_1 \geq \|L_0\|_* + \lambda\|S_0'\|_1 + (1-\beta)(\|\mathcal{P}_{T^\perp} H_L\|_* + \lambda\|P_\Gamma H_L\|_1), \qquad (7.2)$$

where $\beta = \max(\|W\|, \|F\|_\infty)$. Finally, $\|\mathcal{P}_{T^\perp} H_L\|_* + \lambda\|P_\Gamma H_L\|_1$ vanishes if and only if $H_L \in \Gamma^\perp \cap T = \{0\}$.

Lemma 7.2 *Assume that for any matrix M, $\|\mathcal{P}_T\mathcal{P}_{\Gamma^\perp} M\|_F \leq n\|\mathcal{P}_{T^\perp}\mathcal{P}_{\Gamma^\perp} M\|_F$ and take $\lambda > 4/n$. Then (L_0, S_0') is the unique solution if there is a pair (W, F) obeying*

$$UV^* + W + \mathcal{P}_T D = \lambda(sgn(S_0') + F),$$

with $\mathcal{P}_T W = 0$, $\|W\| < 1/2$, $\mathcal{P}_{\Gamma^\perp} F = 0$ and $\|F\|_\infty < 1/2$, and $\|\mathcal{P}_T D\|_F \leq n^{-2}$.

Note that $\|\mathcal{P}_T\mathcal{P}_{\Gamma^\perp} M\|_F \leq n\|\mathcal{P}_{T^\perp}\mathcal{P}_{\Gamma^\perp} M\|_F$ implies $\Gamma^\perp \cap T = \{0\}$, or equivalently $\|\mathcal{P}_{\Gamma^\perp}\mathcal{P}_T\| < 1$. Indeed if $M \in \Gamma^\perp \cap T$, $\mathcal{P}_T\mathcal{P}_{\Gamma^\perp} M = M$ while $\mathcal{P}_{T^\perp}\mathcal{P}_{\Gamma^\perp} M = 0$, and thus $M = 0$.

Proof It follows from (7.2) together with the same argument as in the proof of Lemma 7.2 that

$$\|L_0 + H_L\|_* + \lambda\|S_0' - H_S\|_1 \geq \|L_0\|_* + \lambda\|S_0'\|_1 + \frac{1}{2}\left(\|\mathcal{P}_{T^\perp} H_L\|_* + \lambda\|\mathcal{P}_\Gamma H_L\|_1\right) - \frac{1}{n^2}\|\mathcal{P}_T H_L\|_F.$$

Observe now that

$$\begin{aligned}
\|\mathcal{P}_T H_L\|_F &\leq \|\mathcal{P}_T\mathcal{P}_\Gamma H_L\|_F + \|P_T\mathcal{P}_{\Gamma^\perp} H_L\|_F \\
&\leq \|\mathcal{P}_T\mathcal{P}_\Gamma H_L\|_F + n\|\mathcal{P}_{T^\perp}\mathcal{P}_{\Gamma^\perp} H_L\|_F \\
&\leq \|\mathcal{P}_T\mathcal{P}_\Gamma H_L\|_F + n(\|\mathcal{P}_{T^\perp}\mathcal{P}_\Gamma H_L\|_F + \|\mathcal{P}_{T^\perp} H_L\|_F) \\
&\leq (n+1)\|\mathcal{P}_\Gamma H_L\|_F + n\|P_{T^\perp} H_L\|_F.
\end{aligned}$$

Using both $\|\mathcal{P}_\Gamma H_L\|_F \leq \|\mathcal{P}_\Gamma H_L\|_1$ and $\|\mathcal{P}_{T^\perp} H_L\|_F \leq \|\mathcal{P}_{T^\perp} H_L\|_*$, we obtain

$$\|L_0 + H_L\|_* + \lambda\|S_0' - H_S\|_1 \geq \|L_0\|_* + \lambda\|S_0'\|_1 + \left(\frac{1}{2} - \frac{1}{n}\right)\|\mathcal{P}_{T^\perp} H_L\|_* + \left(\frac{\lambda}{2} - \frac{n+1}{n^2}\right)\|\mathcal{P}_\Gamma H_L\|_1.$$

The claim follows from $\Gamma^\perp \cap T = \{0\}$. $\blacksquare$

Lemma 7.3 *Under the assumptions of Theorem 1.2, the assumption of Lemma 7.2 is satisfied with high probability. That is, $\|\mathcal{P}_T\mathcal{P}_{\Gamma^\perp} M\|_F \leq n\|\mathcal{P}_{T^\perp}\mathcal{P}_{\Gamma^\perp} M\|_F$ for all M.*

Proof Set $\rho_0 = p_0(1-\tau)$ and $M' = \mathcal{P}_{\Gamma^\perp}M$. Since $\Gamma \sim \text{Ber}(\rho_0)$, Theorem 2.6 gives $\|\mathcal{P}_T - \rho_0^{-1}\mathcal{P}_T\mathcal{P}_\Gamma\mathcal{P}_T\| \leq 1/2$ with high probability. Further, because $\|\mathcal{P}_\Gamma\mathcal{P}_T M'\|_F = \|\mathcal{P}_\Gamma\mathcal{P}_{T^\perp}M'\|_F$, we have

$$\|\mathcal{P}_\Gamma\mathcal{P}_T M'\|_F \leq \|\mathcal{P}_{T^\perp}M'\|_F.$$

In the other direction,

$$\begin{aligned}
\rho_0^{-1}\|\mathcal{P}_\Gamma\mathcal{P}_T M'\|_F^2 &= \rho_0^{-1}\langle \mathcal{P}_T M', \mathcal{P}_T\mathcal{P}_\Gamma\mathcal{P}_T M'\rangle \\
&= \langle \mathcal{P}_T M', \mathcal{P}_T M'\rangle + \langle \mathcal{P}_T M', (\rho_0^{-1}\mathcal{P}_T\mathcal{P}_\Gamma\mathcal{P}_T - \mathcal{P}_T)M'\rangle \\
&\geq \|\mathcal{P}_T M'\|_F^2 - \frac{1}{2}\|\mathcal{P}_T M'\|_F^2 = \frac{1}{2}\|\mathcal{P}_T M'\|_F^2.
\end{aligned}$$

In conclusion, $\|\mathcal{P}_{T^\perp}M'\|_F \geq \|\mathcal{P}_\Gamma\mathcal{P}_T M'\|_F \geq \frac{\rho_0}{2}\|\mathcal{P}_T M'\|_F$, and the claim follows since $\frac{\rho_0}{2} \geq \frac{1}{n}$. ∎

Thus far, our analysis shows that to establish our theorem, it suffices to construct a pair (Y^L, W^S) obeying

$$
\begin{cases}
\|\mathcal{P}_{T^\perp}Y^L\| < 1/4, \\
\|\mathcal{P}_T Y^L - UV^*\|_F \leq n^{-2}, \\
\mathcal{P}_{\Gamma^\perp}Y^L = 0, \\
\|\mathcal{P}_\Gamma Y^L\|_\infty < \lambda/4,
\end{cases}
\quad \text{and} \quad
\begin{cases}
\mathcal{P}_T W^S = 0, \\
\|W^S\| \leq 1/4, \\
\mathcal{P}_\Omega W^S = \lambda\text{sgn}(S_0'), \\
\mathcal{P}_{\Omega_{\text{obs}}^\perp}W^S = 0, \\
\|\mathcal{P}_\Gamma W^S\|_\infty \leq \lambda/4.
\end{cases}
\tag{7.3}
$$

Indeed, by definition, $Y^L + W^S$ obeys

$$Y^L + W^S = \lambda(\text{sgn}(S_0') + F),$$

where F is as in Lemma 7.2, and it can also be expressed as

$$Y^L + W^S = UV^* + W + \mathcal{P}_T D,$$

where W and $\mathcal{P}_T D$ are as in this lemma as well.

Construction of the dual certificate Y^L. We use the golfing scheme to construct Y^L. Think of $\Gamma \sim \text{Ber}(\rho_0)$ with $\rho_0 = p_0(1-\tau)$ as $\cup_{1 \leq j \leq j_0}\Gamma_j$, where the sets $\Gamma_j \sim \text{Ber}(q)$ are independent, and q obeys $\rho_0 = 1 - (1-q)^{j_0}$. Here, we take $j_0 = \lceil 3\log n\rceil$, and observe that $q \geq \rho_0/j_0$ as before. Then starting with $Y_0 = 0$, inductively define

$$Y_j = Y_{j-1} + q^{-1}\mathcal{P}_{\Gamma_j}\mathcal{P}_T(UV^* - Y_{j-1}),$$

and set

$$Y^L = Y_{j_0} = q^{-1}\sum_j \mathcal{P}_{\Gamma_j}Z_{j-1}, \quad Z_j = (P_T - q^{-1}\mathcal{P}_T\mathcal{P}_{\Gamma_j}\mathcal{P}_T)Z_{j-1}. \tag{7.4}$$

By construction, $\mathcal{P}_{\Gamma^\perp}Y^L = 0$. Now just as in Section (3.2), because q is sufficiently large, $\|Z_j\| \leq e^{-j}\|UV^*\|_\infty$ and $\|Z_j\|_F \leq e^{-j}\sqrt{r}$, both inequality holding with large probability. The proof is now identical to that in (2.5). First, the same steps show that

$$\|\mathcal{P}_{T^\perp}Y^L\| \leq C\sqrt{\frac{n\log n}{q}}\|UV^*\|_\infty = C'\sqrt{\frac{\mu r(\log n)^2}{n\rho_0}}.$$

Whenever $\rho_0 \geq C_0 \frac{\mu r (\log n)^2}{n}$ for a sufficiently large value of the constant C_0 (which is possible provided that ρ_r in (1.6) is sufficiently small), this terms obeys $\|\mathcal{P}_{T^\perp} Y^L\| \leq 1/4$ as required. Second,

$$\|\mathcal{P}_T Y^L - UV^*\|_F = \|Z_{j_0}\|_F \leq e^{-3 \log n}\sqrt{r} \leq n^{-2}.$$

And third, the same steps give

$$\|Y^L\|_\infty \leq q^{-1}\|UV^*\|_\infty \sum_j e^{-j} \leq 3(1 - e^{-1})\sqrt{\frac{\mu r (\log n)^2}{\rho_0^2 n^2}}.$$

Now it suffices to bound the right-hand side by $\frac{\lambda}{4} = \frac{1}{4}\sqrt{\frac{1-\tau}{n\rho_0}}$. This is automatic when $\rho_0 \geq C_0 \frac{\mu r (\log n)^2}{n}$ whenever C_0 is sufficiently large and, thus, the situation is as before. In conclusion, we have established that Y^L obeys (7.3) with high probability.

Construction of the dual certificate W^S. We first establish that with high probability,

$$\|\mathcal{P}_T \mathcal{P}_\Omega\| \leq \sqrt{\tau' p_0}, \quad \tau' = \tau + \tau_0, \tag{7.5}$$

where $\tau_0(\tau)$ is a continuous function of τ approaching zero when τ approaches zero. In other words, the parameter τ' may become arbitrary small constant by selecting τ small enough. This claim is a straight application of Corollary 2.7. We also have

$$\|\mathcal{P}_\Omega \mathcal{P}_{(T+\Omega_{\text{obs}}^\perp)} \mathcal{P}_\Omega\| \leq 2\tau'. \tag{7.6}$$

with high probability. This second claim uses the identity

$$\mathcal{P}_\Omega \mathcal{P}_{(T+\Omega_{\text{obs}}^\perp)} \mathcal{P}_\Omega = \mathcal{P}_\Omega \mathcal{P}_T (\mathcal{P}_T \mathcal{P}_{\Omega_{\text{obs}}} \mathcal{P}_T)^{-1} \mathcal{P}_T \mathcal{P}_\Omega.$$

This is well defined since the restriction of $\mathcal{P}_T \mathcal{P}_{\Omega_{\text{obs}}} \mathcal{P}_T$ to T is invertible. Indeed, Theorem 2.6 gives $\mathcal{P}_T \mathcal{P}_{\Omega_{\text{obs}}} \mathcal{P}_T \geq \frac{p_0}{2}\mathcal{P}_T$ and, therefore, $\|(\mathcal{P}_T \mathcal{P}_{\Omega_{\text{obs}}} \mathcal{P}_T)^{-1}\| \leq 2p_0^{-1}$. Hence,

$$\|\mathcal{P}_\Omega \mathcal{P}_{(T+\Omega_{\text{obs}}^\perp)} \mathcal{P}_\Omega\| \leq 2p_0^{-1}\|\mathcal{P}_\Omega \mathcal{P}_T\|^2,$$

and (7.6) follows from (7.5).

Setting $E = \text{sgn}(S_0')$, this allows to define W^S via

$$W^S = \lambda(\mathcal{I} - \mathcal{P}_{(T+\Omega_{\text{obs}}^\perp)})(\mathcal{P}_\Omega - \mathcal{P}_\Omega \mathcal{P}_{(T+\Omega_{\text{obs}}^\perp)} \mathcal{P}_\Omega)^{-1} E$$
$$:= (\mathcal{I} - \mathcal{P}_{(T+\Omega_{\text{obs}}^\perp)})(W_0^S + W_1^S),$$

where $W_0^S = \lambda E$, and $W_1^S = \mathcal{R}E$ with $\mathcal{R} = \sum_{k \geq 1}(\mathcal{P}_\Omega \mathcal{P}_{(T+\Omega_{\text{obs}}^\perp)} \mathcal{P}_\Omega)^k$. The operator $\mathcal{R}$ is self-adjoint and obeys $\|\mathcal{R}\| \leq \frac{2\tau'}{1-2\tau'}$ with high probability. By construction, $\mathcal{P}_T W^S = \mathcal{P}_{\Omega_{\text{obs}}^\perp} W^S = 0$ and $\mathcal{P}_\Omega W^S = \lambda \text{sgn}(S_0')$. It remains to check that both events $\|W^S\| \leq 1/4$ and $\|\mathcal{P}_\Gamma W^S\|_\infty \leq \lambda/4$ hold with high probability.

Control of $\|W^S\|$. For the first term, we have $\|(\mathcal{I} - \mathcal{P}_{(T+\Omega_{\text{obs}}^\perp)})W_0^S\| \leq \|W_0^S\| = \lambda\|E\|$. Because the entries of E are i.i.d. and take the value ± 1 each with probability $p_0 \tau/2$, and the value 0 with probability $1 - p_0 \tau$, standard arguments give

$$\|E\| \leq 4\sqrt{np_0(\tau + \tau_0)}$$

35

with large probability. Since $\lambda = 1/\sqrt{p_0 n}$, $\|W_0^S\| \le 4\sqrt{\tau + \tau_0} < 1/8$ with high probability, provided τ is small enough.

For the second term, $\|(\mathcal{I} - \mathcal{P}_{(T+\Omega_{\mathrm{obs}}^{\perp})})W_1^S\| \le \lambda\|\mathcal{R}E\|$, and the same covering argument as before gives

$$\mathbb{P}(\lambda\|\mathcal{R}(E)\| > t) \le 2 \times 6^{2n} \exp\Big(-\frac{t^2}{2\lambda^2\sigma^2}\Big) + \mathbb{P}(\|\mathcal{R}\| \ge \sigma).$$

Since $\lambda = 1/\sqrt{np_0}$ this shows that $\|W^S\| \le 1/4$ with high probability, since one can always choose σ, or equivalently $\tau' = \tau + \tau_0$, sufficiently small.

Control of $\|\mathcal{P}_\Gamma W^S\|_\infty$. For $(i,j) \in \Gamma$, we have

$$W_{ij}^S = \langle e_i e_j^*, W^S \rangle = \lambda \langle X(i,j), E \rangle,$$

where

$$X(i,j) = (\mathcal{P}_\Omega - \mathcal{P}_\Omega \mathcal{P}_{(T+\Omega_{\mathrm{obs}}^{\perp})}\mathcal{P}_\Omega)^{-1}\mathcal{P}_\Omega \mathcal{P}_{(T+\Omega_{\mathrm{obs}}^{\perp})^{\perp}} e_i e_j^*.$$

The same strategy as before gives

$$\mathbb{P}\Big(\sup_{(i,j)\in G} |W_{ij}^S| > \frac{\lambda}{4}\Big) \le 2n^2 \exp\Big(-\frac{1}{8\sigma^2}\Big) + \mathbb{P}\Big(\sup_{(i,j)\in G} \|X(i,j)\|_F > \sigma\Big).$$

It remains to control the Frobenius norm of $X(i,j)$. To do this, we use the identity

$$\mathcal{P}_\Omega \mathcal{P}_{(T+\Omega_{\mathrm{obs}}^{\perp})^{\perp}} e_i e_j^* = \mathcal{P}_\Omega \mathcal{P}_T (\mathcal{P}_T \mathcal{P}_{\Omega_{\mathrm{obs}}} \mathcal{P}_T)^{-1} \mathcal{P}_T e_i e_j^*,$$

which gives

$$\|\mathcal{P}_\Omega \mathcal{P}_{(T+\Omega_{\mathrm{obs}}^{\perp})^{\perp}} e_i e_j^*\|_F \le \sqrt{\frac{4\tau'}{p_0}} \|\mathcal{P}_T e_i e_j^*\|_F \le \sqrt{\frac{8\mu r\tau'}{np_0}}$$

with high probability. This follows from the fact that $\|(\mathcal{P}_T \mathcal{P}_{\Omega_{\mathrm{obs}}} \mathcal{P}_T)^{-1}\| \le 2p_0^{-1}$ and $\|\mathcal{P}_\Omega \mathcal{P}_T\| \le \sqrt{p_0\tau'}$ as we have already seen. Since we also have $\|(\mathcal{P}_\Omega - \mathcal{P}_\Omega \mathcal{P}_{(T+\Omega_{\mathrm{obs}}^{\perp})}\mathcal{P}_\Omega)^{-1}\| \le \frac{1}{1-2\tau'}$ with high probability,

$$\sup_{(i,j)\in\Gamma} \|X(i,j)\|_F \le \frac{1}{1-2\tau'}\sqrt{\frac{8\mu r\tau'}{np_0}}.$$

This shows that $\|\mathcal{P}_\Gamma W^S\|_\infty \le \lambda/4$ if τ', or equivalently τ, is sufficiently small.

Acknowledgements

References

[1] R. Basri and D. Jacobs. Lambertian reflectance and linear subspaces. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 25(2):218–233, 2003.

[2] A. Beck and M. Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM Journal on Imaging Sciences*, 2(1):183–202, Mar 2009.

[3] S. Becker, J. Bobin, and E. J. Candès. NESTA: A fast and accuract first-order method for sparse recovery. *preprint*, 2009.

[4] M. Belkin and P. Niyogi. Laplacian eigenmaps for dimensionality reduction and data representation. *Neural Computation*, 15(6):1373–1396, 2003.

[5] D.P. Bertsekas. *Constrained Optimization and Lagrange Multiplier Method*. Academic Press, 1982.

[6] J. Cai, E. J. Candès, and Z. Shen. A singular value thresholding algorithm for matrix completion. preprint, 2008.

[7] E. J. Candès and Y. Plan. Matrix completion with noise. *Proceedings of the IEEE* (to appear), 2009.

[8] E. J. Candès and B. Recht. Exact matrix completion via convex optimzation. *Found. of Comput. Math.*, 9:717–772, 2009.

[9] E. J. Candès, J. Romberg, and T. Tao. Robust uncertainty principles: exact signal reconstruction from highly incomplete frequency information. *IEEE Trans. Inform. Theory*, 52(2):489–509, 2006.

[10] E. J. Candès and T. Tao. The power of convex relaxation: Near-optimal matrix completion. *IEEE Trans. Inf. Theory* (to appear), 2009.

[11] V. Cevher, A. Sankaranarayanan, M. Duarte, D. Reddy, R. Baraniuk, and R. Chellappa. Compressive sensing for background subtraction. In *Proceedings of European Conference on Computer Vision (ECCV)*, 2009.

[12] V. Chandrasekaran, S. Sanghavi, P. Parrilo, and A. Willsky. Rank-sparsity incoherence for matrix decomposition. *preprint*, 2009.

[13] S. Chen, D. Donoho, and M. Saunders. Atomic decomposition by basis pursuit. *SIAM Review*, 43(1):129–159, 2001.

[14] S. Dewester, S. Dumains, T. Landauer, G. Furnas, and R. Harshman. Indexing by latent semantic analysis. *Journal of the Society for Information Science*, 41(6):391–407, 1990.

[15] C. Eckart and G. Young. The approximation of one matrix by another of lower rank. *Psychometrika*, 1:211–218, 1936.

[16] M. Fazel, H. Hindi, and S. Boyd. Log-det heuristic for matrix rank minimization with applications to Hankel and Euclidean distance matrices. In *Proceedings of the American Control Conference*, pages 2156–2162, Jun 2003.

[17] M. Fischler and R. Bolles. Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, 24:381–385, 1981.

[18] A. Georghiades, P. Belhumeur, and D. Kriegman. From few to many: Illumination cone models for face recognition under variable lighting and pose. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 23(6), 2001.

[19] R. Gnanadesikan and J. Kettenring. Robust estimates, residuals, and outlier detection with multiresponse data. *Biometrics*, 28:81–124, 1972.

[20] D. Goldfarb and S. Ma. Convergence of fixed point continuation algorithms for matrix rank minimization. *preprint*, 2009.

[21] M. Grant and S. Boyd. CVX: Matlab software for disciplined convex programming (web page and software). http://stanford.edu/∼boyd/cvx, June 2009.

[22] D. Gross. Recovering low-rank matrices from few coefficients in any basis. *CoRR*, abs/0910.1879, 2009.

[23] D. Gross, Y-K. Liu, S. T. Flammia, S. Becker, and J. Eisert. Quantum state tomography via compressed sensing. *CoRR*, abs/0909.3304, 2009.

[24] T. Hey, S. Tansley, and K. Tolle. *The Fourth Paradigm: Data-Intensive Scientific Discovery*. Microsoft Research, 2009.

[25] H. Hotelling. Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology*, 24:417–441, 1933.

[26] P. Huber. *Robust Statistics*. Wiley and Sons, 1981.

[27] I. Jolliffe. *Principal Component Analysis*. Springer-Verlag, 1986.

[28] Q. Ke and T. Kanade. Robust ℓ^1-norm factorization in the presence of outliers and missing data. In *Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition*, 2005.

[29] R. Keshavan, A. Montanari, and S. Oh. Matrix completion from a few entries. 2009.

[30] S. Kontogiorgis and R. Meyer. A variable-penalty alternating direction method for convex optimization. *Mathematical Programming*, 83:29–53, 1989.

[31] M. Ledoux. *The Concentration of Measure Phenomenon*. American Mathematical Society, 2001.

[32] L. Li, W. Huang, I. Gu, and Q. Tian. Statistical modeling of complex backgrounds for foreground object detection. *IEEE Transactions on Image Processing*, 13(11):1459–1472, 2004.

[33] Z. Lin, M. Chen, L. Wu, and Y. Ma. The augmented Lagrange multiplier method for exact recovery of a corrupted low-rank matrices. *Mathematical Programming*, submitted, 2009.

[34] Z. Lin, A. Ganesh, J. Wright, L. Wu, M. Chen, and Y. Ma. Fast convex optimization algorithms for exact recovery of a corrupted low-rank matrix. In *Computational Advances in Multi-Sensor Adaptive Processing (CAMSAP)*, 2009.

[35] P. Lions and B. Mercier. Splitting algorithms for the sum of two nonlinear operators. *SIAM Journal on Numerical Analysis*, 16(6):964–979, 1979.

[36] Z. Liu and L. Vandenberge. Interior-point method for nuclear norm approximation with application to system identification. *SIAM Journal on Matrix Analysis and Applications*, 31(3):1235–1256, 2009.

[37] S. Ma, D. Goldfarb, and L. Chen. Fixed point and Bregman iterative methods for matrix rank minimization. *preprint*, 2009.

[38] Y. Nesterov. A method of solving a convex programming problem with convergence rate $O(1/k^2)$. *Soviet Mathematics Doklady*, 27(2):372–376, 1983.

[39] Y. Nesterov. Smooth minimization of non-smooth functions. *Mathematical Programming*, 103(1), 2005.

[40] Y. Nesterov. Gradient methods for minimizing composite objective functions. *Technical Report - CORE - Universite Catholique de Louvain*, 2007.

[41] Netflix, Inc. The Netflix prize. *http://www.netflixprize.com/*.

[42] S. Osher, M. Burger, D. Goldfarb, J. Xu, and W. Yin. An iterative regularization method for total variation-based image restoration. *Multiscale Modeling and Simulation*, 4:460–489, 2005.

[43] C. Papadimitriou, P. Rghavan, H. Tamaki, and S. Vempala. Latent semantic indexing, a probabilistic analysis. *Journal of Computer and System Sciences*, 61(2):217–235, 2000.

[44] B. Recht. A simpler approach to matrix completion. *CoRR*, abs/0910.0651, 2009.

[45] B. Recht, M. Fazel, and P. Parillo. Guaranteed minimum rank solution of matrix equations via nuclear norm minimization. submitted to *SIAM Review*, 2008.

[46] J. Tenenbaum, V. de Silva, and J. Langford. A global geometric framework for nonlinear dimensionality reduction. *Science*, 290(5500):2319–2323, 2000.

[47] F. De La Torre and M. Black. A framework for robust subspace learning. *International Journal on Computer Vision*, 54:117–142, 2003.

[48] R. Vershynin. Math 280 lecture notes. Available at `http://www-stat.stanford.edu/~dneedell/280.html`, 2007.

[49] W. Yin, E. Hale, and Y. Zhang. Fixed-point continuation for ℓ^1-minimization: Methodology and convergence. *preprint*, 2008.

[50] W. Yin, S. Osher, D. Goldfarb, and J. Darbon. Bregman iterative algorithms for ℓ_1-minimization with applications to compressed sensing. *SIAM Journal on Imaging Sciences*, 1(1):143–168, 2008.

[51] X. Yuan and J. Yang. Sparse and low-rank matrix decomposition via alternating direction methods. *preprint*, 2009.

[52] Z. Zhou, A. Wagner, H. Mobahi, J. Wright, and Y. Ma. Face recognition with contiguous occlusion using Markov random fields. In *Proceedings of International Conference on Computer Vision (ICCV)*, 2009.